AMERICAN MATHEMATICAL SOCIETY

Lecture Notes Prepared in Connection With the

Summer Seminar on Space Mathematics

held at

Cornell University, Ithaca, New York

July 1, 1963 - August 9, 1963

Part IV

UNPUBLISHED PRELIMINARY DATA

Supported by the

National Aeronautics and Space Administration under Research Grant NsG 358

Air Force Office of Scientific Research under Grant AF-AFOSR 258-63

Army Research Office (Durham) under Contract DA-31-124-ARO(D)-82

Atomic Energy Commission under Contract AT(30-1)-3164

Office of Naval Research under Contract Nonr(G)00025-63

National Science Foundation under NSF Grant GE-2234

Special Computation Procedures for Differential Equations

by S.V. Parter

I. Introduction.

Many problems of applied mathematics arise naturally as differential equations. In most cases there is no hope of finding an explicit, closed representation of the solution. Thus we are led to the computer. However, the availability of high-speed computers does not mean that "practical men" can give up the analytical study of differential equations.

Indeed, in some sense, the great advance in our computational ability requires that we put more effort into the analytical study. After all, twenty years ago we could only shrug our shoulders at these problems. Now we can and do attempt to get approximate results. And, in order to get computational results that are meaningful, we must do some analysis.

In these lectures, I hope to present some of the ideas and results in this area.

II. Ordinary Differential Equations.

The simplest problem is Pure Initial-Value Problem

$$(2.1) \qquad y' = f(x, y), \qquad y(x_0) = y_0.$$

Here, $y = y(x)$ may be a vector and then (2.1) represents a system of equations. It is well known that almost every Initial-Value problem may be put in this form. For, suppose we start with

$$(2.2) \qquad y^{(\nu)} = f(x, y, y^1, \ldots, y^{\nu-1}), \; y^{(j)}(x_0) = y_0^j,$$

$$j = 0, \; 1, \ldots, \nu-1.$$

Then we set $Z_1 = y_1$, $z_2 = y^{(1)}$, ..., $z_1 = y^{(\nu-1)}$, and we write (2.2) as

$$(2.1')\quad \begin{cases} Z_1' = Z_2 \\ Z_2' = Z_3 \\ \cdots \\ Z_{\nu-1}' = Z_\nu \\ Z_\nu' = f(x, z, t_1, \ldots, z_\nu), \; Z_j(x_0) = y_0^{j-1}, \\ \qquad\qquad\qquad\qquad j = 1, 2, \ldots, \nu. \end{cases}$$

We now turn to the question of numerical methods for approximating the solution $y(x)$ of (2.1).

We assume that $f(x, y)$ is continuous in (x, y) and satisfies a Lipschitz condition in y, i.e., there is constant L such that

$$(2.3)\qquad \|f(x, y) - f(x, z)\| \leq L\|y - z\|.$$

These conditions assure us of the existence of a unique solution $y(x)$ in some neighborhood of x_0. For those of you who are skeptical of such mathematical niceties, let us consider two examples.

Example 1: $y' = y^{1/2}$, $y(0) = 0$.

Then $y_1(x) = 0$ and $y_2(x) = \frac{1}{4}x^2$ are solutions in the

interval $0 \leq x \leq 1$.

Example 2: $y' = 1 + y^2$, $y(0) = 0$.

In this case $y(x) = \tan x$ and there is no solution in the "larger" interval $0 \leq x \leq \pi$.

Now, let an increment h be chosen; then we seek the values y_k which approximate $y(kh + x_0)$.

The simplest formula we can use is

$$(2.4) \qquad Y_{k+1} = Y_k + hf(x_k, Y_k), \; Y_0 = y_0.$$

Here $x_k = x_0 + kh$. This is an example of a Single-Step Method which we write as

$$(2.5) \qquad Y_{k+1} = Y_k + h\phi(x_k, Y_k; h).$$

From the form of (2.5) one might think that

$$\phi(x_k, Y_k; h) = f(x_k, Y_k)$$

is the only "natural" choice. However, let me point out that the familiar Runge-Kutta method is also of this form.

Theorem 1: Let $f(x, y)$ be continuous in (x, y) and satisfy the Lipschitz condition (2.3). Moreover, let $\phi(x, y; h)$ also satisfy a Lipschitz condition. Then

$$(2.6) \qquad \lim_{h \to 0+} \phi(x, y, h) = f(x, y)$$

is a necessary and sufficient condition for the convergence of the solution $\{Y_k\}$ of (2.5) to $y(\bar{x})$ in the limit as

$$h \longrightarrow 0,\ kh + x_0 = \bar{x}.$$

Theorem 2: Let $f(x, y)$ be continuous in (x, y) and satisfy the Lipschitz condition (2.3); let $\phi(x, y; h)$ also satisfy a Lipschitz condition. Moreover, let the "consistency" condition (2.6) be satisfied. Finally, let the truncation error be $O(h^p)$, $p > 0$, i.e., if $y(x)$ is the solution of (2.1), then

$$y\left[(k+1)\ h\right] = y(kh) + h\left[\phi(x_k,\ y(kh);\ h) + O(h^p)\right].$$

Then, as $h \longrightarrow 0$ and $kh = \bar{x}$, we have

$$\|Y_k - y(kh)\| = O(h^p). \tag{2.7}$$

We will omit the proof of Theorem 1, as it is technically complicated. However, let us give a proof of Theorem 2.

Proof of Theorem 2. Let

$$E_k = Y_k - y(kh).$$

Then, from (2.5) and (2.7) we have

$$E_{k+1} = E_k + h\left\{\left[\phi(x_k,\ Y_k;\ h) - \phi(x_k,\ y(kh);\ h\right] + O(h^p)\right\}.$$

Therefore, since ϕ also satisfies a Lipschitz condition,

$$\|E_{k+1}\| \leq \|E_k\| + hL\left\{\|E_k\| + O(h^p)\right\},$$

i.e.,

$$\|E_{k+1}\| \leq (1 + h\lfloor) \|E_k\| + M h^{p+1},$$

or

$$\|E_{k+1}\| \leq (1 + h\lfloor)^2 \|E_{k-1}\| + [(1 + h\lfloor) + 1] M h^{p+1},$$

or

$$\|E_{k+1}\| \leq (1 + h\lfloor)^{k+1} \|E_0\| + \left[\sum_{j=0}^{k} (1 + h\lfloor)^j\right] M h^{p+1}.$$

We sum the geometric progression and find that

$$\|E_{k+1}\| \leq (1 + h\lfloor)^{k+1} \cdot \|E_0\| + \frac{(1+h\lfloor)^{k+1} - 1}{1 + h\lfloor - 1} M h^{p+1}.$$

That is

$$\|E_{k+1}\| \leq e^{h\lfloor (k+1)} \left[\|E_0\| + \frac{M}{\lfloor} h^p\right].$$

If we now assume that $E_0 = 0$, we obtain the desired result.

Having these two theorems, we will leave the topic of single-step methods for the initial-value problem.

Of course, there are other methods of treating the initial-value problem. Let us consider the Linear Multi-Step Methods. The simplest such method is

$$(2.8) \qquad Y_{k+1} = Y_{k-1} + 2hf(x_k, Y_k).$$

Notice that in this case we must specify both Y_0 and Y_1. Now Y_0 can be taken as y_0, but it is almost impossible to specify Y_1 exactly.

We now consider only the scaler case, i.e., $y(x)$ is a scaler, not a vector.

In general, we have constants $\alpha_1, \alpha_2, \ldots, \alpha_k$, $\beta_1, \ldots, \beta_k$, and we use the recurrence relation

$$(2.9)\qquad \alpha_k Y_{n+k} + \alpha_{k-1} Y_{n+k-1} + \ldots + \alpha_0 Y_n = h\left\{\beta_k f(x_{k+n}, Y_{k+n}) + \ldots + \beta_0 f(x_n, Y_n)\right\},$$

or

$$(2.9a)\qquad \sum_{j=0}^{k} \alpha_j Y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, Y_{n+j}).$$

Of course, we assume $\alpha_k \neq 0$. If $\beta_k = 0$, we say that (2.9) is an "explicit" linear multi-step method. On the other hand, if $\beta_k \neq 0$, then we have an "implicit" method.

Example 3: Consider the linear multi-step method

$$(2.10)\qquad Y_{n+3} + \tfrac{3}{2}Y_{n+2} - 3Y_{n+1} + \tfrac{1}{2}Y_n = 3h\, f(x_{n+2}, Y_{n+2}).$$

One can easily verify that -- provided $f(x, y)$ is nice enough --

$$y(x_{n+3}) + \tfrac{3}{2}y(x_{n+2}) - 3y(x_{n+1}) + \tfrac{1}{2}y(x_n) = 3h\, f(x_n, y(x_n)) + O(h^4).$$

That is, (2.10) is a consistent approximation to (2.1) and with a small truncation error.

When I present this example in class, I ask my students to try the two problems

$$y' = -y, \qquad y(0) = 1$$
$$y' = y, \qquad y(0) = 1$$

in the range $0 \leq x \leq 1$ with $h = 0.01$. For those of you who have access to a computer, I recommend these problems.

You will find them very instructive.

In any case, a simple analysis -- but one which is too lengthy to give here -- shows that the solutions of (2.10) are unstable and do not converge to the solution of (1.2) Moreover, this is true even in the simplest cases.

The results in this case are too complicated to prove in the short space of time we have here. However, they are easy enough to state. (See Henrici [1] for details.)

Let

$$(2.11) \qquad \rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j.$$

We have a Stability Condition: For all ζ which are roots of $\rho(\zeta) = 0$, we must have

$$(2.11a) \qquad |\zeta| \leq 1.$$

Moreover, if ζ is a double root of $\rho(\zeta) = 0$, we must have

$$(2.11b) \qquad |\zeta| < 1$$

And, as before, we have a Consistency Condition: This condition -- in words -- merely says that the solutions of (2.1), i.e., the solutions of the differential equation "almost" satisfy the difference equation (2.9). It is rather easy to verify that a necessary condition for consistency is

$$(2.12)\quad \begin{cases} \sum_{j=0}^{k} \alpha_j = 0, \\ \sum_{j=0}^{k} (j\alpha_j - \beta_j) = 0 \end{cases}$$

Definition: A linear multi-step method given by two sets of coefficients $\{\alpha_j\}$, $\{\beta_j\}$ is called convergent if the error $E_n = |Y_n - y(x_n)| \to 0$ as $h \to 0$ and $n \to \infty$ in such a way that $x_n \equiv \bar{x}$, provided only $E_j \to 0$ for $j = 0, 1, \ldots, k-1$ and the function $f(x, y)$ is continuous in (x, y) for all y and $|x-x_0| < b$ (for some $b > 0$) and also satisfies a Lipschitz condition in y.

Theorem 3: The linear multi-step method given by the two sets of coefficients $\{\alpha_j\}$, $\{\beta_j\}$ is convergent if and only if both the stability condition (2.11a), (2.11b) and the consistency condition (2.12) are satisfied.

Theorem 4: Suppose the linear multi-step method (2.9) is convergent. Let $y(x)$ be a solution of (2.1). Assume also that

$$\sum_{j=0}^{k} \alpha_j y(x_{n+j}) = h \sum \beta_j f(x_{n+j}, y(x_{n+j})) + O(h^{p+1}).$$

Then

$$|E_m| = O(h^p).$$

Now, let us mention another approach to our basic problem. This approach is motivated by the fact that many good linear multi-step methods are implicit, i.e., $\beta_k \neq 0$. Therefore the solution of (2.9) becomes messy. So we consider Predictor-Corrector Methods of the form

$$Y^{*}_{n+k} + \sum_{j=0}^{k-1} a_j Y_{n+j} = h \sum_{j=0}^{k-1} b_j f(x_{n+j}, Y_{n+j})$$

(2.13)

$$\sum_{j=0}^{k} \alpha_j Y_{n+j} = h \left\{ \rho_k f(x_{n+k}, Y^{2}_{n+k}) + \sum_{j=0}^{k-1} \beta_j f_j \right\}, \alpha_k \neq 0.$$

The general idea here is to use a high-order predictor formula and a "stable" correcter formula.

Before we leave these initial-value problems, a few remarks are in order.

The motivation for linear multi-step methods is clearly the desire to use more accurate formulae. However, one should note that these methods can lead to many complications. First of all, one must have accurate methods for more "starting" values than are implied by the problem. Also, there is the problem of stability. Finally, there is a whole host of problems associated with the slow decay of certain components of the error which have been introduced by the linear multi-step method itself. Once more, let me recommend the book by Henrici.

Now, let us say a few words about "boundary-value" problems. Consider the problem

$$(2.14) \quad \begin{cases} -(py')' = f(x), & 0 \leq x \leq 1 \\ y(0) = y(1) = 0, \end{cases}$$

where the function $p(x) \geq p_0 > 0$ is a "smooth" function.

We take $h = 1/M$, where M is an integer. Once more, let Y_j represent an approximation to $y(jh)$, and let

$p_{j+1/2} = p\left[(j+1/2)h\right]$. The first set of equations to come to mind are

$$(2.15)\quad \begin{cases} Y_0 = Y_m = 0 \\ -p_{j-1/2}\, Y_{j-1} + (p_{j-1/2} + p_{j+1/2})\, Y_j - p_{j+1/2}\, Y_{j+1} = \\ \qquad h^2\, f(x_j). \end{cases}$$

Now we have two problems:

(1.) Can we solve these equations?

(2.) Assuming the answer to (1.) is yes, does the error $E_k = \left| Y_k - y(kh) \right| \to 0$?

In both cases the answer is yes! Let us look at the first question.

Consider a general tridiagonal system of linear equations of the form $Y_0 = Y_n = 0$

$$a_j\, Y_{j-1} + b_j\, Y_j + C_j\, Y_{j+1} = Q_j$$

where

$$(2.16)\quad b_j \geq \left| a_j \right| + \left| C_j \right|.$$

Then, when we look at the straight-forward elimination procedure, we discover the following algorithm. Let

$$(2.17)\quad G_0 = F_0 = 0$$

$$(2.17a)\quad D_k = b_k + a_k\, G_{k-1} \text{ for } k = 1, 2, \ldots, M-1.$$

$$(2.17b)\quad G_k = -C_k/D_k$$
$$F_k = (Q_k - A_k\, F_{k-1})/D_k.$$

Then

$$(2.18a)\quad Y_{M-1} = F_{M-1}$$

and for $j < M-1$, we have

$$Y_j = G_j\, Y_{j+1} + F_j. \tag{2.18b}$$

Thus, our equation can be solved rather easily. Moreover, the condition (2.16) guarantees that this procedure is numerically stable.

As for the second question, if we multiply (2.15) by Y_j and sum on j, we have

$$\sum_{(j)} \left[Y_j \cdot p_{j-1/2} (Y_j - Y_{j-1}) + Y_j \cdot p_{j+1/2} (Y_j - Y_{j+1}) \right] =$$

$$\sum_{(j)} h^2 f_j Y_j$$

or

$$\sum_j p_{j-1/2} \left[\frac{Y_j - Y_{j-1}}{h} \right]^2 = \sum f_j Y_j$$

or

$$h \sum_j \left[\frac{Y_j - Y_{j-1}}{h} \right]^2 \leq \frac{1}{p_0} \sqrt{h \sum |f_j|^2} \cdot \sqrt{h \sum |Y_j|^2}. \tag{2.19}$$

And, for any set Z_j ; $j = 0, 1, 2, \ldots, M$ with $Z_0 = Z_m = 0$, we have

$$h \sum_j |Z_j|^2 \leq \frac{h^2}{2(1-\cos \pi h)}\, h \sum \left[\frac{Z_j - Z_{j-1}}{h} \right]^2.$$

This last result is easily established by elementary matrix theory. Since

$$\frac{h^2}{2(1-\cos \pi h)} \longrightarrow \frac{1}{\pi^2} \text{ as } h \longrightarrow 0$$

we can claim the existence of a constant $K > 0$ so that

$$\left[h \sum_{(j)} |Z_j|^2 \right] \leq K^2 \left[h \sum_{(j)} \left(\frac{Z_j - Z_{j-1}}{h} \right)^2 \right].$$

Thus, using (2.19), we have

$$\sqrt{h \sum |Y_j|^2} \cdot \sqrt{h \sum \left(\frac{Y_j - Y_{j-1}}{h} \right)^2} \leq \frac{k}{p_0} \sqrt{h \sum |f_j|^2} \cdot \sqrt{h \sum |Y_j|^2}$$

i.e.

$$\sqrt{h \sum \left(\frac{Y_j - Y_{j-1}}{h} \right)^2} \leq \frac{k}{p_0} \sqrt{h \sum |f_j|^2}. \tag{2.20}$$

Let $k > r$. Then

$$Y_k - Y_r = h \sum_{j=r+1}^{k} \left[\frac{Y_j - Y_{j-1}}{h} \right].$$

Therefore,

$$|Y_k - Y_r| \leq h \sum_{j=r+1}^{k} \left| \frac{Y_j - Y_{j-1}}{h} \right| \leq \sqrt{(k-r)h} \cdot \sqrt{h \sum_{(j)} \left(\frac{Y_j - Y_{j-1}}{h} \right)^2}.$$

That is, using (2.20),

$$|Y(kh) - Y(rh)| \leq |kh - rh|^{1/2} \cdot \frac{K}{p_0} \sqrt{h \sum |f_j|^2}. \tag{2.21a}$$

And if $r = 0$,

$$|Y_k| \leq \frac{K}{p_0} \sqrt{h \sum |f_j|^2}. \tag{2.21b}$$

It is now an easy matter to prove the convergence of the $\{Y_k\}$ to the solution of the boundary-value problem. The simplest approach is merely to observe that the error E_k satisfies a similar difference equation. However, in

this case. the right-hand-side $f_j \longrightarrow 0$ as $h \longrightarrow 0$. Hence

$$h \sum \left| f_j \right|^2 \longrightarrow 0 \text{ as } h \longrightarrow 0$$

and the convergence follows from (2.21b).

III. Partial Differential Equations.

Once more, let us consider the Initial-Value Problem. Consider the special case of a first-order-linear system of the form

$$(3.1) \qquad \begin{cases} \dfrac{\partial U}{\partial t} = P(x, t; D)\, U, \\ U(x, 0) = U_0(x). \end{cases}$$

Here, $x = (x_1, x_2, \ldots, x_n)$ and U is a vector $\{ U_1, U_2, \ldots, U_N \}$ and $P(x, t; D)$ is a matrix polynomial in the $(\frac{\partial}{\partial x_\cdot})$ with coefficients depending on (x, t).

Let's look at a very simple special case --

$$(3.2) \qquad \begin{cases} \dfrac{\partial U}{\partial t} = -\dfrac{\partial U}{\partial x} \\ U(x, 0) = f(x) \end{cases}$$

One can easily prove that the solution to this problem is

$$(3.2a) \qquad U(x, t) = f(x - t).$$

Indeed, to verify that (3.2a) is a solution (assuming that $f(x)$ is differentiable) is an exercise in Calculus.

Even though we know the solution of this problem, let us look at some finite-difference approximations.

$$(A.) \qquad \frac{v(x, t+k) - v(x, t)}{k} = \frac{v(x+h, t) - v(x, t)}{h},$$

which reduces to

$$(3.3) \qquad v(x, t+k) = (1+\lambda)\, v(x, t) - \lambda\, v(x+h, t)$$

where $\lambda = k/h$. Repeated application of (3.3) leads to a formula of the form

$$(3.4) \qquad v(x, nk) = \sum_{j=0}^{n} a_j\, v(x+jh, 0) = \sum_{j=0}^{n} a_j\, f(x+jh).$$

The exact values of the coefficients aj are inessential for our present argument. The important fact is that the value of $v(x, t)$ depends only on the value of $f(x)$ at points to the <u>right</u> of x. On the other hand, from (3.2a), we see that the solution of the differential equation depends on a value of $f(x)$ at a point to the <u>left</u> of x; namely $(x-t)$. It is now an easy matter to construct an initial function $f(x)$ which is very smooth -- say infinitely differentiable -- and the solutions of (3.3) cannot possibly converge to $f(x-t)$. For example, let $f(x) > 0$ for $x < 0$ and $f(x) = 0$ for $x \geq 0$. Then, we see from (3.4) that $v(x, nk) = 0$ for all $x > 0$. On the other hand, $U(x, nk) > 0$ for all $x < nk$.

All right, let's try another approach --

$$(B.) \qquad \frac{v(x, t+k) - v(x, t)}{k} = \frac{v(x, t) - v(x-h, t)}{h},$$

which reduces to

$$(3.5) \qquad v(x, t + k) = (1 - \lambda)\, v(x, t) + \lambda\, v(x-h, t).$$

In this case, an argument very similar to the one we have just given shows that we must take $\lambda \leq 1$.

Thus, these examples illustrate the general situation. As is the case of ordinary differential equations, it is not enough to have a consistent approximation to the differential equation. Moreover, the restrictions on the difference schemes are frequently restrictions on ratios of the step-lengths in the different coordinate directions.

Let us look at another example, the heat equation

$$(3.6) \qquad \frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \qquad U(x, 0) = f(x).$$

We try the difference scheme

$$\frac{v(x, t + k) - v(x, t)}{k} = \frac{1}{h^2}\left\{ v(x - h, t) - 2v(x, t) + v(x + h, t) \right\}.$$

In this case, the necessary condition is

$$(3.7) \qquad \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}.$$

This result is particularly interesting because, unlike our earlier results, there does not seem to be any obvious relationship between (3.7) and the analytical properties of the solution of (3.6). That statement is not strictly

true for, if (3.7) is satisfied, then

$$(3.8) \qquad \operatorname*{Sup}_x \left| v(x, nk) \right| \leq \operatorname*{Sup}_x \left| v(x, 0) \right|$$

and the physically interesting solutions of (3.6) satisfy a similar estimate. On the other hand, as we shall see, there are convergent difference schemes for the heat equation which do not enjoy property (3.8). Let me put it this way: Since (3.7) implies (3.8), it is easy to prove that the solutions $v(x, nk)$ of the difference scheme converge to $U(x, t)$ the solution of (3.6) provided that (3.7) holds. However, it is not apparent that (3.7) is a necessary condition for convergence.

Let us return to our general problem. If we select an $h = (h_1, h_2, \ldots, h_n)$ and k then a finite difference equation should give us approximations to the solution $U(x, t)$ at the lattice points $(j_1 h_1, j_2 h_2, \ldots, j_n h_n, rk)$. Let $v(r)$ denote the vector

$$\left\{ v(j_1 h_1, \ldots, j_n h_n, rk) \right\}_{j_1, j_2 \ldots, j_n = -\infty}^{\infty},$$

and let $\| v(r) \|$ denote some norm on these "vectors". For example, we could have

$$(3.9a) \qquad \| v(r) \| = \operatorname*{Sup}_{j_s h_s} \left| v(j_1 h_1, j_2 h_2, \ldots, j_n h_n, rk) \right|$$

or

$$(3.9b) \qquad \| v(r) \| = h_1 h_2 \cdots h_n \cdot \sum_{j_s} \left| v(j_1 h_1, j_2 h_2, \ldots, j_n h_n, rk) \right|^2$$

etc.

If B is a linear operator (i.e., an infinite matrix in this case) acting on these vectors, we define

$$\|B\| = \operatorname{Sup}_{x=10} \frac{\|Bx\|}{\|x\|} . \tag{3.10}$$

Suppose we have a finite-difference approximation to (3.1) of the form

$$\begin{cases} \mathcal{V}(r+1) = B(r) \cdot \mathcal{V}(r) \\ \mathcal{V}(j_1h_1, j_2h_2, \ldots, j_nh_n, 0) = U_0(j_1h_1, j_2h_2, \ldots, j_nh_n). \end{cases} \tag{3.11}$$

In (3.11) we should write $\mathcal{V}(r+1; h)$ and $B(r; h)$ since these operators and vectors will depend on the lattice h and the increment k, etc.

Definition: We say that the family of operators $\{B(r; h)\}$ is stable in the interval $0 \leq t \leq T$ and in the $\|\ \|$ norm if there is a constant K, depending on T, such that

$$\|B(r; h)\, B(r-1; h) \ldots B(j+1; h)\, B(j; h)\| \leq K \tag{3.12}$$

for all r, j with

$$0 \leq j \leq r \leq T/k. \tag{3.12a}$$

The basic convergence argument is based on this notion and a simple argument which we saw earlier in our discussion of single-step methods for ordinary differential equations.

Theorem: Suppose (3.1) has a solution u(x, t). Let w(r)

be the "vector" determined by $u(x, rk)$, i.e.

$$w(j_1h_1, \ldots, j_nh_n, rk) = u(j_1h_1, \ldots, j_nh_n, rk).$$

Assume that (3.11) is a "consistent" approximation to (3.1), i.e.

(3.13) $$w(r+1) = B(r)\, w(r) + a(r)$$

and

(3.14) $$\|a(r)\| = O(k^{1+p}) \qquad p > 0.$$

Finally, assume that the family $\{B(r; h)\}$ is stable for $0 \leq t \leq T$ in the $\|\ \ \|$ norm.

Then, for all $0 \leq t \leq T < \infty$ of the form $t = rk$, we have

(3.15) $$\|w(r) - v(r)\| = O(Tk^p).$$

Proof: Let $E(r) = w(r) - v(r)$. Then, from (3.13) we have

$$\begin{aligned} E(r+1) &= B(r)\, E(r) + a(r) \\ &= B(r)\, B(r-1)\, E(r-1) + B(r)\, a(r-1) + a(r) \\ &= B(r)\, B(r-1) \ldots B(0) \cdot E(0) + \\ &\qquad \sum_{j=2}^{r} \left[B(r)\, B(r-1) \ldots B(j)\right] A(j-1) + A(r). \end{aligned}$$

Since the family $\{B(r; h)\}$ is stable, and $E(0) = 0$, we have

$$\|E(r)\| \leq K \cdot r \cdot O(k^{1+p}) = K(rk)\, O(k^p).$$

that is

$$\|E(r)\| \leq KT \cdot O(k^p).$$

Thus we have shown that, under reasonable conditions,

stability implies convergence. A natural question is -- what about the converse? In the appropriate theoretical setup, the answer is that stability is in fact also necessary for convergence. Let me refer you to the excellent book by Richtmyer [4]. Of course, as a practical matter, stability is absolutely essential!

In general, there is no obvious way to determine the stability or instability of a difference scheme. However, in some cases we can get a grip on these ideas. And, these precise results lead to relatively good rules of thumb.

Consider the case where $P(x, t; D)$ has constant coefficients. That is

$$(3.16) \qquad \frac{\partial u}{\partial t} = P(D)\, u$$

$$= \sum_{\ell=0}^{R} \sum_{\ell_1+\ell_2+\ldots+\ell_n=\ell} A_{\ell_1\ell_2\cdots\ell_n} \frac{\partial^{\ell}}{\partial x_{i_1}^{\ell_1}\ldots x_n^{\ell_n}} U$$

where the $A_{\ell_1\cdots\ell_n}$ are constant matrices.

Moreover, let us assume that the difference equation (3.11) takes the form

$$(3.17) \qquad v(r+1) = B\, v(r)$$

where B is a fixed operation. More specifically, we assume

$$(3.18)\qquad v(j_1h_1, \ldots, j_nh_n;\ rk+k) =$$

$$\sum_{|\ell_1| + |\ell_2| \cdots |\ell_n| \leq R_1} B_{\ell_1 \ell_2 \cdots \ell_n}$$

$$v\left[(j_1+\ell_1)\, h_1, \ \ldots, \ (j_n+\ell_n)\, h_n;\ rk\right]$$

where the $B_{\ell_1, \ \ldots, \ \ell_n}$ are constant matrices. Consider the matrix-valued function

$$(3.19)\qquad \beta(\theta_1, \theta_2, \ldots, \theta_n) = \sum B_{\ell_1 \ell_2 \ldots \ell_n}\ e^{i \sum \ell_j \theta_j}.$$

A rather straightforward application of Fourier analysis, which can be done in several ways, leads to the following conclusion: Let the norm be chosen as in (3.9b). Then

$$\left\| B^r \right\| = \underset{|\theta_j| \leq \pi}{\text{MAX}} \left\| \beta^r(\theta_1, \theta_2, \ldots, \theta_n) \right\|_F$$

where $\| \quad \|_F$ represents the finite-dimensional matrix norm of $\beta^r(\theta_1, \theta_2, \ldots, \theta_n)$.

Thus, our problem has been reduced to a finite-dimensional problem. This problem is still not trivial. In fact, it is sometimes rather messy. However, we do have a method of analysis.

Let us return to our earlier examples. Consider the equations (3.2) and the difference equations (A.) and (B.). In case (A.), we use equation (3.3) and find that

$$\beta(\theta) = (1+\lambda) - \lambda e^{i\theta}.$$

and

$$\beta(\overline{\Pi}) = (1+2\lambda)$$

$$\left\|\beta^{r}(\overline{\Pi})\right\| = (1+2\lambda)^{r} \longrightarrow \infty \text{ as } r \longrightarrow \infty, \; rk \leq T$$

if λ is a constant. Thus the nonconvergent method (A.) is unstable.

In case (B.) we find that

$$\beta(\theta) = (1-\lambda) + \lambda e^{-i\theta}$$

$$\left|\beta(\theta)\right|^{2} = \left[1 - \lambda(1-\cos\theta)\right]^{2} + \lambda^{2}\sin^{2}\theta$$

$$= 1 - 2\lambda(1-\cos\theta) + \lambda^{2}(1-2\cos\theta + \cos^{2}\theta) + \lambda^{2}\sin^{2}\theta$$

$$= 1 + 2\lambda(\lambda-1)(1-\cos\theta)$$

Since $1-\cos\theta \geq 0$, $\text{MAX}\left|\beta(\theta)\right| \leq 1$ if and only if $\lambda \leq 1$. And, in this case, $\text{MAX}\left|\beta^{r}(\theta)\right| = \text{MAX}\left|\beta(\theta)\right|^{r}$.

Turning now to the heat equation (3.6) and the related difference equation, we find that

$$\beta(\theta) = 1 + 2\frac{\Delta t}{(\Delta x)^{2}}(1-\cos\theta).$$

Thus $\beta(\theta)$ is real, $\beta(\theta) \leq 1$, and $\beta(\theta) \geq -1$ if and only if (3.7) holds.

Well, this is a fine analysis. But what about the general problem of differential equations and a fortiori

difference equations with variable coefficients. In the general case, we have the following rule of thumb: For each value (x_0, t_0), $0 \leq t_0 \leq T$, consider the differential equation and difference equations with all coefficients evaluated at (x_0, t_0). These difference equations are of the form we have analyzed. And, if for (x_0, t_0) the corresponding difference equations are "stable", then they are also stable in the variable coefficient case.

The validity of the above rule of thumb has not been established in complete generality. However, there are some fairly general results justifying this procedure.

Before proceeding, let us point out that if, with a finite difference equation of the form (3.11) which we write for short

$$v(r+1) = B \cdot v(r),$$

we associate the norm

$$\|v(r)\|^2 = h_1 \cdot h_2 \cdot h_3, \ldots, h_n \sum |v(\ldots, r)|^2,$$

then

$$\|B\| = \text{Sup} \frac{\|BU\|}{\|U\|} = \max \|\beta(\theta_1, \ldots, \theta_n)\|_F, \tag{3.20}$$

where $\| \quad \|_F$ is as before.

Let us now consider as a further example the wave equation

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial^2 U}{\partial x^2}. \tag{3.21}$$

We try the difference scheme

$$\frac{v_k^{n+1} - 2v_k^n + v_k^{n-1}}{(\Delta t)^2} = \frac{v_{k-1}^n - 2v_k^n + v_{k+1}^n}{(\Delta x)^2},$$

where we write $v(k\Delta x, n\Delta t) = v_k^n$. We transform the difference equation to

$$(3.22) \qquad v_k^{n+1} = 2v_k^n - v_k^{n-1} + \lambda^2(v_{k-1}^n - 2v_k^n + v_{k+1}^n),$$

where we have put $\lambda = \Delta t/\Delta x$. We could use a geometric argument to establish stability criteria, since we already know that the solution of (3.21) is $U = f(x+t) + g(x-t)$. The domain of dependence argument tells us that for stability we must have $\lambda \leq 1$. It is important to recognize, however, that (3.22) is not of the form (3.11). The time dependence is on the nth and the n-1th values.

To avoid this difficulty, we write (3.22) as the system

$$(3.23) \qquad w_k^{n-1} = v_k^n$$

$$v_k^{n+1} = \lambda^2(v_{k-1}^n + v_{k+1}^n) + 2(1-\lambda^2)v_k^n - w_k^n.$$

To compute $\beta(\theta)$, (3.19), write the β-matrix

$$(3.24) \qquad \beta = \begin{bmatrix} 0 & 1 \\ -1 & 2(1-\lambda^2) + 2\lambda^2\cos\theta \end{bmatrix}.$$

It's possible to show that $\|\beta^r(\theta)\| \leq K$, but we will not do it. Instead we will prove the necessary condition for

the truth of this condition, i.e., the eigenvalues are less than or equal to one in absolute value. The eigenvalues μ must satisfy

$$(3.25) \qquad \mu^2 - 2\left[1-\lambda^2(1-\cos\theta)\right]\mu + 1 = 0.$$

It follows immediately that the product of μ_+ and μ_- must equal 1; if the roots are real, then they are either +1 or -1, or one is larger than the other in absolute value. Complex roots can only be one in magnitude. This leads to the condition

$$\left\{1 - \lambda^2(1-\cos\theta)\right\}^2 \leq 1,$$

or

$$(3.26) \qquad -1 \leq 1 - \lambda^2(1-\cos\theta) \leq 1,$$

which is true if and only if $\lambda \leq 1$.

By another approach, assume the solution of (3.22) is of the form

$$(3.27) \qquad v_k^n = \mu^n e^{i(k\Delta x)\theta} = \mu^n e^{ik\theta} \qquad (\Delta x = 1).$$

Substituting into (3.22) gives

$$\mu^{n+1} - 2\mu^n + \mu^{n-1} = \lambda^2\mu^n\left[e^{i\theta} - 2 + e^{i\theta}\right].$$

Factoring μ^n gives

$$(3.28) \qquad \mu^2 - 2\left[1 + \lambda^2(\cos\theta-1)\right]\mu + 1 = 0,$$

which is identical to (3.25).

IV. Practical Problems in Partial Differential Equations.

Consider the heat equation (3.6), but ask that it be satisfied in

$$0 \leq x \leq 1, \qquad t > 0,$$

with conditions given

$$\begin{aligned} &U(x, 0) = f(x) && 0 \leq x \leq 1 \\ (4.1)\qquad &U(0, t) = g(t) \\ &U(1, t) = h(t). && t > 0 \end{aligned}$$

In the notation of (3.22), write the following family of difference equations:

$$\frac{v_k^{n+1} - v_k^n}{\Delta t} = \alpha \frac{v_{k-1}^n - 2v_k^n + v_{k+1}^n}{(\Delta x)^2} + (1-\alpha)\frac{v_{k-1}^{n+1} - 2v_k^{n+1} + v_{k+1}^{n+1}}{(\Delta x)^2},$$

where $\lambda \equiv \Delta t/(\Delta x)^2$ and $0 \leq \alpha \leq 1$. Rewriting, we obtain

$$(4.2)\qquad -\lambda(1-\alpha)v_{k-1}^{n+1} + \left[1 + 2(1-\alpha)\lambda\right]v_k^{n+1} - \lambda(1-\alpha)v_{k+1}^{n+1} =$$

$$\alpha\lambda\left[v_{k-1}^n - 2v_k^n + v_{k+1}^n\right] + v_k^n,$$

which, when applied to a system of mesh points, yields a tri-diagonal system of linear equations in terms of the known boundary conditions, which can be solved for any α. Is the system stable?

As before, assume the solution of (4.2) is

$$v_k^n = \beta^n e^{ik\theta}.$$

We then obtain

$$(4.3) \qquad \beta = \frac{1 + 2\alpha\lambda(\cos\theta - 1)}{1 + 2(1-\alpha)\lambda(1-\cos\theta)}.$$

It's clear that $\beta \leq 1$ for all α, λ, θ. To meet the stability condition $\beta \leq -1$, it is necessary and sufficient that

$$(4.4) \qquad (2\alpha - 1)\lambda \leq 1/2$$

Note that if $\alpha \leq 1/2$, (4.4) is no restriction on λ, since λ is always ≥ 0; if $\alpha > 1/2$, on the other hand, λ is restricted.

To study the stability and convergence of (4.2), you must make a detailed study of the tri-diagonal matrix in an equation for the error, and this gives the same result as (4.4).

We now have a whole family of finite difference equations in terms of one parameter to solve (3.6), and in particular, if we take $\alpha < 1/2$, the equations are unconditionally stable.

To prove the convergence of the method for $\alpha = 0$, consider

$$(4.5) \qquad v_k^{n+1} - v_k^n = \lambda \left[v_{k-1}^{n+1} - 2v_k^{n+1} + v_{k+1}^{n+1} \right]$$

and study $\operatorname{Max}_k v_k^{n+1}$. Then $k = 0$ or $M = 1/\Delta x$, i.e.,

$\underset{k}{\text{Max}}\ v_k^{n+1}$ occurs on the boundary, or $v_k^n \geq v_k^{n+1}$. Suppose $\underset{k}{\text{Max}}\ v_k^{n+1}$ is not on the boundary. At the interior point where v_k^{n+1} is a maximum,

$$v_{k-1}^{n+1} - 2v_k^{n+1} + v_{k+1}^{n-1} \leq 0.$$

Then, since $\lambda > 0$, $v_k^{n+1} \leq v_k^n$. Similarly, $\underset{k}{\text{Min}}\ v_k^{n+1}$ is true for $k = 0$ or $M = 1/\Delta x$ or $v_k^n \leq v_k^{n+1}$. From these two statements together, it follows that

$$(4.6) \qquad \underset{\substack{0 \leq n \leq N \\ k}}{\text{Max}} \left| v_k^n \right| \leq \text{Max} \left| f \right|, \left| g \right|, \left| h \right|$$

for $0 \leq t \leq n\Delta t$

$0 \leq x \leq 1$,

which is the stability condition.

To see that this insures convergence, consider the error equation

$$(4.7) \qquad E_k^{n+1} - E_k^n = \lambda \left[E_{k-1}^{n+1} - 2E_k^{n+1} + E_{k+1}^{n+1} \right] + \sigma_k^n,$$

where $\left| \sigma_k^n \right| \leq L(\Delta t)^{1+p}$ and $E_k^0 = 0$, $E_0^n = E_M^n = 0$. Let $\underset{k}{\text{Max}}\ E_k^{n+1} = E_j^{n+1}$. If $j \neq 0$, $j \neq M$, then

$$(4.8) \qquad \left| E_j^{n+1} \right| \leq \left| E_j^n \right| + \left| \sigma_k^n \right|$$

by arguments similar to those used to prove (4.6). Now it is straightforward to show by continued inequalities that

$$(4.9) \qquad \max_k \left|E_k^{n+1}\right| \leq \max_k \left|E_k^n\right| + L\Delta t^{1+p},$$

and thus convergence is assured.

As an example of a two-dimensional problem, consider the heat equation

$$(4.10) \qquad \frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}$$

for $t > 0$, and (x, y) in some region R. The initial and boundary conditions are written in a form analagous to (4.1).

If we write $v_{k,j}^n$ for $v(k\Delta x, j\Delta y, n\Delta t)$, we get the difference equation

$$(4.11) \qquad \frac{v_{k,j}^{n+1} - v_{k,j}^{n-1}}{\Delta t} = \frac{v_{k-1,j} - 2v_{k,j} + v_{k+1,j}}{(\Delta x)^2} + \frac{v_{k,j-1} - 2v_{k,j} + v_{k,j+1}}{(\Delta y)^2}.$$

The superscripts on the right-hand side have deliberately been omitted. If we put the superscript n+1 on each term, we could prove a Min-Max principle similar to (4.6) and the corresponding convergence results, but the resulting system of equations, when applied to a problem, are untenable. To avoid this, consider the case of (4.11) written as

$$(4.12) \qquad \frac{v_{k,j}^{n+2} - v_{k,j}^{n+1}}{\Delta t} = \frac{v_{k-1,j}^{n+1} - 2v_{k,j}^{n+1} + v_{k-1,j}^{n+1}}{(\Delta x)^2} + \frac{v_{k,j-1}^{n+2} - 2v_{k,j}^{n+2} + v_{k,j+1}^{n+2}}{(\Delta y)^2},$$

where the superscript changes are obvious. This gives a tri-diagonal system which must be inverted for each value of k. Note carefully the difference in (4.11) written as

$$(4.13)\qquad \frac{v_{k,j}^{n+1} - v_{k,j}^{n-1}}{\Delta t} = \frac{v_{k-1,j}^{n+1} - 2v_{k,j}^{n+1} + v_{k+1,j}^{n+1}}{(\Delta x)^2} + \frac{v_{k,j-1}^{n} - 2v_{k,j}^{n} + v_{k,j+1}^{n}}{(\Delta y)^2}.$$

The use of (4.12), combined with (4.13), forms the well known alternating direction method, which is stable and therefore converges, provided Δt is given the same value for every pair of steps. The proof of stability is similar to that used for (4.2).

To illustrate a practical problem involving an elliptic equation, a boundary value problem, consider

$$(4.14)\qquad \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f(x, y) \text{ in a region R}$$

$$\text{with } U = g(x, y) \text{ on the boundary of R.}$$

As a first step, we must pick Δx and Δy, i.e., set up a basic lattice in R.

To accomodate the boundary by means of the basic mesh, we can either modify the boundary to fit the mesh, i.e., use only mesh points which are within the boundary and estimate the influence of the boundary on the nearest interior mesh point by interpolation, for example; or we can modify our difference equation for interior points to

account for different mesh lengths near the boundary. Let us assume that we take the former method. Then we have the following formulation: If $(k\Delta x, j\Delta y)$ is an interior point,

$$(4.15) \qquad \Delta_h^v \equiv \frac{v_{k+1,j} - 2v_{k,j} + v_{k-1,j}}{(\Delta x)^2} +$$

$$\frac{v_{k,j-1} - 2v_{k,j} + v_{k,j+1}}{(\Delta y)^2} = f_{k,j}.$$

If $k\Delta x, j\Delta y$ is a boundary point, then $v_{k,j}$ is known.

Most of the effort in research involving (4.15) is in finding methods for solving the resulting linear systems. However, it is easy to prove convergence of (4.15) as follows:

Observe that if $f_{k,j} \geqslant 0$ for all k,j, then the maximum of $v_{k,j}$ is assumed on the boundary. The proof is by contradiction. Suppose the maximum is interior. Then, if it is at the point k,j, the left-hand side of (4.15) is less than zero. But $f(x, y) \geqslant 0$. Then (4.15) is satisfied only if both sides are zero. Extending consideration to the boundary points and noting that (4.15) expresses the fact that $v_{k,j}$ is an average of its four neighboring points, it is clear that the maximum of $v_{k,j}$ cannot occur in the interior of R without a contradiction.

Similar results follow for $f_{k,j} \leqslant 0$ in R. We conclude that Min $v_{k,j}$ must occur on the boundary. These results already imply the existence of a solution of the equations

(4.15). We have a linear system of equations in as many unknowns for which either there is always a solution, or there exists a nontrivial solution of the homogeneous equation. But there is no such nontrivial solution, in view of the above results.

If the function $f(x, y)$ changes sign, let

$$w_{k,j} = \frac{(k\Delta x)^2 + (j\Delta y)^2}{4}, \tag{4.16}$$

which is defined everywhere. We find that

$$\Delta_h w = 1. \tag{4.17}$$

Let $F = \text{Max} \left| f_{k,j} \right|$ and v be a solution of $\Delta_h v = f$. Then

$$\Delta_h(wF - u) \geqslant 0, \tag{4.18}$$

which implies that Max $(wF - v)$ occurs on the boundary. But $\text{Max}\ (v_{k,j}) \leq (r^2/4)F + \underset{\text{bdy}}{\text{Max}} \left|v\right|$, where r is at least the radius of the smallest circle which encloses R. Similarly, we can show that $\Delta_h(v - wF) \leq 0$, and hence $\text{Max} \left|v_{k,j}\right| \leq Cr^2F + \underset{\text{bdy}}{\text{Max}} \left|v\right|$, and convergence is assured.

V. Inversion of a Matrix.

In general we have large matrices to invert. How do we invert them? Let us write a typical matrix in a particular structural form. This will be significant in what follows. Write the matrix $V = (v_{ij})$ as a vector:

(5.1)
$$V = \begin{pmatrix} V_1 \\ V_2 \\ \cdot \\ \cdot \\ \cdot \\ V_n \end{pmatrix} \quad \text{where } V_j = \begin{pmatrix} v_{1j} \\ v_{2j} \\ \cdot \\ \cdot \\ \cdot \\ v_{\ell j} \end{pmatrix},$$

i.e., we arrange everything by lines. V_j is the vector of unknowns in the jth line of any of the typical linear systems discussed above. Then the problem takes the form

(5.2)
$$AV = k$$

where A is a matrix, V is the vector of unknowns (5.1), and k is a vector of known values arising from the boundary conditions.

The matrix A has the following structure:

(5.3)
$$A = \begin{bmatrix} D_1 & F_1 & & & 0 \\ E_2 & D_2 & F_2 & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & F_{m-1} \\ & 0 & & E_m & D_m \end{bmatrix} = D + E + F.$$

It is a block tri-diagonal matrix, which can be written as the sum of three matrices, D, E, and F, as in (5.3). The matrices D_j are again tri-diagonal, and therefore are particularly easy to invert.

An iterative method for finding the solution of a

problem written in the form (5.2) is to write

$$A = P - N,$$

assuming the P^{-1} is easily inverted and A is nonsingular. (5.2) becomes

$$PV = NV + k, \tag{5.4}$$

and we obtain an iterative equation by placing superscripts as shown:

$$PV^{(\nu+1)} = NV^{(\nu)} + k. \tag{5.5}$$

If we define the error $E^{(\nu)} = V^{(\nu)} - V$, (5.4) and (5.5) give

$$E^{(\nu+1)} = (P^{-1}N)^{\nu+1}E^{(0)}. \tag{5.6}$$

From this it follows that the iterative method will be convergent if Max $|\lambda| < 1$ for all λ which are eigenvalues of $P^{-1}N$, which is equivalent to: For all λ such that $\det\{\lambda P - N\} = 0$.

The value of λ, if either of these conditions is satisfied, enables us to estimate the "cost" of alternative methods of iteration. Suppose we have two iteration schemes: a) $A = P_0 - N_0$ with error $E_0^{(\nu)}$ and $\lambda_0 =$ Max $|$eigenvalue of $P_0^{-1}N_0|$ and b) $A = P_1 - N_1$ with error $E_1^{(\nu)}$ and $\lambda_1 = $ Max $|$ eigenvalue of $P_1^{-1}N_1|$. Then it can be shown that

$$\|E_k^{(\nu+1)}\| \sim \lambda_k^{\nu+1} \|E_k^{(0)}\| \qquad k = 0 \text{ or } 1. \tag{5.7}$$

Then, taking the logarithm of (5.7),

$$(5.8) \qquad \frac{\log\left(\dfrac{\|E_k^{(\nu+1)}\|}{\|E_k^{(0)}\|}\right)}{\log \lambda_k} \sim \nu + 1,$$

which tells us how many iterations it would take to accomplish a fixed-ratio decrease in the norm of the error. It is therefore important to be able to estimate the value of λ.

To apply iteration to our problem, write $P = D$ and $N = -(E + F)$, and we obtain what is known as the Jacobi block iteration method:

$$(5.9) \qquad DV^{(\nu+1)} = -(E + F)V^{(\nu)} + k.$$

As an alternative but closely related method, consider the Successive Over-Relaxation method -- S.O.R. Here

$$P = \frac{1}{\omega}(D + \omega E)$$

$$N = \frac{1}{\omega}\left[(1-\omega)\ D - \omega F\right],$$

where ω is some real number. This gives the iteration scheme

$$(5.10) \qquad (D + \omega E)V^{(\nu+1)} = \left[(1-\omega)\ D + \omega F\right]V^{(\nu)} + \omega k.$$

The major problem in each of the formulae (5.10) and (5.9) is the inversion of the D_j.

There is a relationship between the eigenvalues associated with each of the above methods which is shown in the following:

Theorem: Let λ = Max | eigenvalue of Jacobi method | and let μ = Max | eigenvalue of S.O.R. method |. Then

$$(\mu + \omega - 1)^2 = \lambda^2\omega^2\mu.$$

Proof: λ arises as a root of

$$(5.11) \qquad \begin{vmatrix} \lambda D_1 & F_1 & & & 0 \\ E_2 & \lambda D_2 & F_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & F_{m-1} \\ 0 & & & E_m & \lambda D_m \end{vmatrix} = 0,$$

and μ is a root of

$$(5.12) \qquad \begin{vmatrix} \frac{\mu+\omega-1}{\omega} D_1 & F_1 & & & 0 \\ \mu E_2 & \frac{\mu+\omega-1}{\omega} D_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & F_{m-1} \\ 0 & & & \mu E_m & \frac{\mu+\omega-1}{\omega} D_m \end{vmatrix} = 0.$$

Forming the matrix

$$T(\alpha) = \begin{bmatrix} 1 & & & 0 \\ & \alpha & & \\ & & \alpha^2 & \\ & & & \ddots \\ 0 & & & \alpha^{m-1} \end{bmatrix}$$

and computing $\left| T_{(\alpha)} \right| \cdot Q \cdot \left| T^{-1}_{(\alpha)} \right|$, where Q is the determinant (5.12), we obtain

$$\begin{vmatrix} \frac{\mu+\omega-1}{\omega} D_1 & \alpha^{-1}F_1 & & 0 \\ \mu\alpha E_2 & \frac{\mu+\omega-1}{\omega} D_2 & \alpha^{-1}F_2 & \\ & \ddots & \ddots & \ddots \\ & & & \alpha^{-1}F_{m-1} \\ 0 & & \mu\alpha E_m & \frac{\mu+\omega-1}{\omega} D_m \end{vmatrix},$$

which we call Q_0.

$$\text{Now } \left| \alpha I \right| Q_0 = \begin{vmatrix} \alpha\left(\frac{\mu+\omega-1}{\omega}\right) D_1 & F_1 & & \\ \alpha^2\mu E_m & \ddots & \ddots & \\ & \ddots & \ddots & F_{m-1} \\ & & \alpha^2\mu E_m & \frac{\alpha(\mu+\omega-1)}{\omega} D_m \end{vmatrix}.$$

If we put $\alpha^2\mu = 1$ and compare this to (5.11), we have the result stated in the Theorem. There are two interesting cases to consider in the S.O.R. Method.

Case (i). $\omega = 1$. Then $\mu = \lambda^2$, and the S.O.R. method is seen to be superior to Jacobi because a) there is a reduced computer storage requirement, and b) this method is approximately twice as fast as Jacobi, in view of (5.8).

Case (ii). ω = optimum value = ω_b and is such that $1 \leq \omega_b \leq 2$. This gives $\mu_{opt} < 1$ in general.

Now for the finite difference equation for the Laplacian operator, A is positive definite, and D is positive definite. Thus the Jacobi method eigenvalue

$$\lambda = \max_X \frac{|((E + F)X, X)|}{(DX, X)},$$

or, in terms of the results for general iterative methods,

$$\lambda = \max_X \frac{|(NX, X)|}{(PX, X)} = \max_X \frac{|(NX, X)|}{(AX, X) + (NX, X)} < 1,$$

and, since we know $(NX, X) > 0$, the method is always convergent.

To conclude, let us consider the alternating direction method, discussed earlier, as applied to the elliptic equation.

Let us define the matrices H and V such that

$$(H\mu)_{k,j} = -\left\{u_{k-1,j} - 2u_{k,j} + u_{k+1,j}\right\}$$

$$(Vu)_{k,j} = -\left\{u_{k,j-1} - 2u_{k,j} + u_{k,j+1}\right\}.$$

Then equation (4.14) can be written in the form

(5.13) $(\theta_x H + \theta_y V)X = K,$

where $\theta_x = \dfrac{\Delta y^2}{2(\Delta x^2 + \Delta y^2)}, \qquad \theta_y = \dfrac{\Delta x^2}{2(\Delta x^2 + \Delta y^2)},$

and X is a vector of unknowns. Writing (5.13) as $(H_1 + V_1)X = K$, we see our aim is to invert the matrix $H_1 + V_1$, a positive definite matrix, and H_1 and V_1 are positive definite themselves. To do this, put

$$(H_1 + rI)U^{m+1/2} = (rI - V_1)U^m + K$$

$$(V_1 + rI)U^{m+1/2} = (rI - H_1)U^{m+1/2} + K,$$

with $r > 0$. The true solution satisfies both of these equations for any value of r. In fact, we could change r after any two cycles. The error satisfies

$$E^{m+1} = (V_1 + rI)^{-1}\,(rI - H_1)\,(H_1 + rI)^{-1}\cdot(rI - V_1)E^m \equiv W_m E^m.$$

The dominant eigenvalue of W_m can be estimated in two particular cases.

Case I. If the region R is a rectangle, H and V commute, and thus every eigenvalue of W_m is of the form

$$\frac{(r - v_k)}{(v_k - r)} \cdot \frac{(r - h_j)}{(h_j + r)},$$

which is always less than or equal to 1, since each factor is less than one.

Case II. If R is not a rectangle, we must fix the

value of r, which fixes the value of W. Then

$$(rI + V_1)W(rI + V_1)^{-1} = W_0$$

has the same eigenvalues as W, but

$$W_0 = \left[(rI - H_1)\ (rI + H_1)^{-1}\right]\left[(rI - V_1)\ (rI + V_1)^{-1}\right].$$

The terms in both brackets commute separately, and thus we conclude, as above: The method is convergent.

Special Computation Procedures.

Bibliography.

1. Henrici, Peter: Discrete Variable Methods in Ordinary Differential Equations. John Wiley & Sons, Inc. (1962)
2. Anonymous: Modern Computing Methods. Notes Appl. Sci. No. 16, National Physical Laboratory, London (1957)
3. Fox, L.: The Numerical Solution of Two-Point Boundary Problems in Ordinary Differential Equations. Oxford University Press. (1957)
4. Richtmyer, R.D.: Difference Methods for Initial-Value Problems. Interscience, New York. (1957)
5. Varga, R.S.: Matrix Iterative Analysis. Prentice-Hall. Englewood Cliffs, New Jersey. (1962)
6. Douglas, J. Jr.: Survey of Numerical Methods for Parabolic Differential Equations, in "Advances in Computers", Vol. 11, Academic Press. New York, New York. (1961)
7. Young, D.M. and Frank, T.G.: A Survey of Computer Methods for Solving Elliptic and Parabolic Partial Differential Equations, in "I.C.C. Bulletin", Vol. 2, No. 1, January, 1963.
8. Bennett, A.A., W.E. Milne, and H. Bateman: Numerical Integration of Differential Equations, Dover, New York. (1956)

Qualitative Methods in the n - Body Problem

by

Professor H. Pollard

Introduction

The n-body problem is generally concerned with the motion of masses $m_1, \ldots, m_n$ $(n>1)$, moving in inertial space under the attraction of their gravitional forces. In the case of a particle m_j being acted upon by mass m_k, we illustrate the geometry in Figure 1.

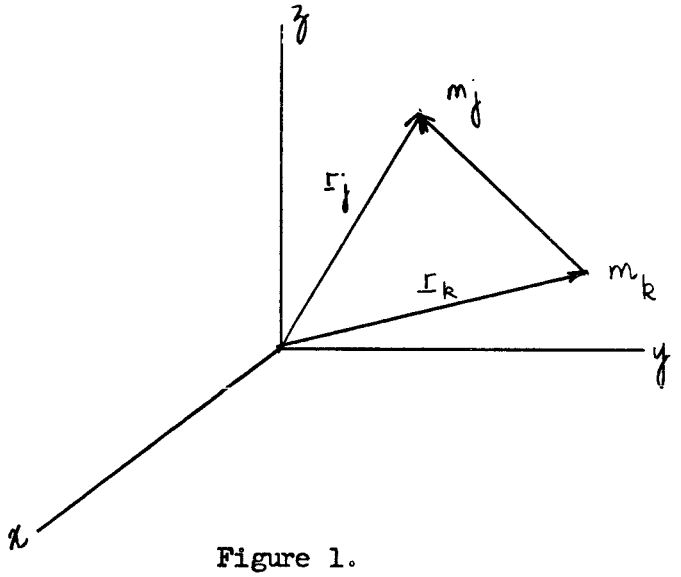


Figure 1.

With position vectors $\underline{r}_j$ and $\underline{r}_k$, the differential equation of motion due to the force on the j^{th} particle by the k^{th} mass is

$$m_k \ddot{\underline{r}}_k = \sum_{\substack{j=1 \\ j\neq k}}^{n} \frac{m_j m_k}{r_{jk}^2} \frac{\underline{r}_j - \underline{r}_k}{r_{jk}} \quad , \quad (k = 1, \ldots, n) \qquad (1)$$

assuming the initial position and velocity are given, i.e., $\underline{r}_k(0)$, $\underline{v}_k(0)$ and $r_{jk}>0$, we seek a solution of (1). To realize what constitutes a solution to a differntial equation, recall the problem

$$\frac{dy}{dx} = f(x, y) \ ,$$

where we seek a solution such that $y_0 = g(x_0)$ for a predetermined point and in general $y = g(x, c)$. Obviously to find c we solve $y_0 = g(x_0, c)$.

However, in actuality we solve $g(x, y, c) = 0$, with $g(x_0, y_0, c) = 0$. Solving for c we find an implicit solution relating x and y. For example, consider

$$\frac{dy}{dx} = \frac{2x + ye^{xy} \cos e^{xy}}{x\, e^{xy} \sin^{xy} + 1} \quad ,$$

with solution

$$x^2 + \sin e^{xy} + y = c.$$

The latter equation is a solution in the sense that if it is differentiated you get the former. Actually such a solution serves no useful purpose unless there exists some transparency that makes it more useful.

Further, assume there exists a set of differential equations

$$\frac{dx}{dt} = f(x, y) \ , \quad \frac{dy}{dt} = g(x, y)$$

with initial conditions x(0) and y(0) given. The problem is to find solutions $x = x(t)$ and $y = y(t)$ satisfying the differential equations and the initial conditions. Simple division of these equations eliminates the variable t and yields $\frac{dy}{dx} = h(x, y)$ where $y = g(x)$. Here we have managed to reduce the system by one, and there is a chance that if the solution is transparent the reduction is useful. Thus mathematicians were led to look for integrals to systems of differential equations. Returning to equation (1), the idea is to reduce it to a system of first order differential equations of the form

$$\frac{dx_k}{dt} = f_k(x, \ldots, x_n) \ , \quad \text{where } k = 1, \ldots, m \ , \quad x_k = x_k(t) \qquad (2)$$

and $x_k(0)$ given for $k = 1, \ldots, m$. The order of (2) is $6n$, with $m = 6n$.

Assume $f(x_1, \ldots, x_m, t)$ is an integral of the system if every solution of the system gives

$$f\left[x_1(t), \ldots, x_m(t), t\right] = \text{constant}, \tag{3}$$

where the constant is determined by

$$f\left[x_1(t), \ldots, x_m(t), t\right] = f\left[\text{initial values}\right] .$$

(3) is an implicit solution of (2) in the sense that if

$$f_k\left[x_1(t), \ldots, x_m(t), t\right] = f\left[\text{initial values}\right] , \quad k = 1, \ldots, m,$$

there exist m equations in n unknowns for which we can solve $x_k = x_k$ (t, initial conditions), and the problem is solved in terms of t and the initial conditions.

Illustrative Central Force Problem

Consider the 2-body problem, n = 2, 12 integrals, with masses moving in a field subject to the inverse square law.

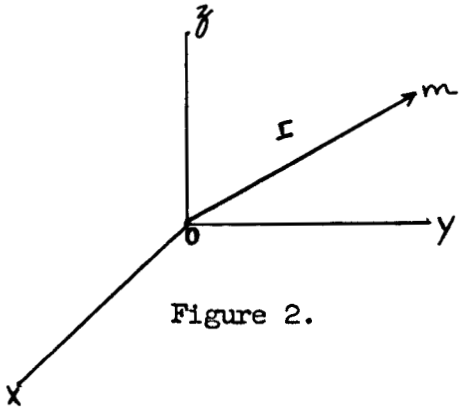


Figure 2.

$$\underline{m\ddot{r}} = -\frac{\mu m}{r^2}\frac{\underline{r}}{r} \quad , \text{ or}$$

$$\underline{\ddot{r}} = -\frac{\mu}{r^3}\underline{r} \ . \tag{4}$$

Note: $\left|\underline{\dot{r}}\right| \neq \dot{r}$

$\left|\underline{r}\right| = r$.

Using Laplace's method, (4) becomes

$$\frac{d}{dt}\frac{\underline{r}}{r} = \frac{r\underline{v} - \dot{r}\underline{r}}{r^2} \quad . \tag{5}$$

Recalling that $r^2 = \underline{r}^2$ and $r\dot{r} = \underline{r} \cdot \dot{\underline{r}}$, (5) can be written as

$$\frac{d}{dt}\frac{\underline{r}}{r} = \frac{(\underline{r} \times \underline{v}) \times \underline{r}}{r^3} = \frac{(\underline{r} \times \underline{v}) \times \ddot{\underline{r}}}{\mu} \quad . \tag{6}$$

Since $\underline{r} \times \ddot{\underline{r}} = 0$ and $\underline{r} \times \underline{v} = \underline{h}$, we have

$$\frac{d}{dt}\frac{\underline{r}}{r} = \frac{-\underline{h} \times \ddot{\underline{r}}}{\mu} \qquad \text{(3 integrals)} \tag{7}$$

Integrating the extremes of (6) and (7), we have

$$\frac{\underline{r}}{r} = \frac{-\underline{h} \times \underline{v}}{\mu} - \underline{e} \qquad \text{or}$$

$$\frac{\underline{r}}{r} + \underline{e} = \frac{\underline{v} \times \underline{h}}{\mu} \qquad \begin{matrix}\text{(3 integrals)} \\ (\underline{e} \text{ has 3 components}).\end{matrix} \tag{8}$$

However, the problem is not complete since

$\frac{dx_k}{dt} = f_k(x_1, \ldots, x_n)$ has solutions of the form

$$f_k(x_1, \ldots, x_n, t) = \text{constant}, \theta$$

and there exists at least one function in which t appears explicitly.

But the 6 integrals in (7) and (8) contain no such function, implying $\underline{e}$ and $\underline{h}$ are not independent of each other. That is,

$$\underline{e} \cdot \underline{h} = 0 \Longrightarrow e_1h_1 + e_2h_2 + e_3h_3 = 0.$$

Thus, in fact, (8) yields but 2 integrals, with the sixth, the time of perihelion passage, still missing.

Returning to the original problem of 2 bodies, with neither body at the origin, we have 12 integrals for the system

$$m_1\ddot{\underline{r}}_1 = \frac{m_1m_2}{r_{12}^2}\frac{\underline{r}_1 - \underline{r}_2}{r_{12}} \quad . \tag{9}$$

$$m_2\ddot{\underline{r}}_2 = \frac{m_1m_2}{r_{12}^2}\frac{\underline{r}_2 - \underline{r}_1}{r_{12}} \quad . \tag{10}$$

Adding (9) and (10) yields

$$m_1\ddot{\underline{r}}_1 + m_2\ddot{\underline{r}}_2 = 0 \quad . \tag{11}$$

Define $\underline{r}_c$ (center of mass) $= \frac{1}{M}(m_1\underline{r}_1 + m_2\underline{r}_2)$, and $M = m_1 + m_2$, so that $\ddot{\underline{r}}_c = 0$. This last equation indicates the center of mass is not accelerating. Integrating to get the velocity of the center of mass,

$\underline{v}_c = \underline{\ell}$ (3 integrals, conservation of linear momentum)

$\underline{r}_c = \underline{\ell}t + \underline{k}$ (3 integrals)

Multiplying (9) by $x\underline{r}_1$ and (10) by $x\underline{r}_2$ and adding, we get

$$m_1 (\underline{r}_1 \times \ddot{\underline{r}}_1) + m_2 (\underline{r}_2 \times \ddot{\underline{r}}_2) = 0.$$

Integration yields

$$m_1(\underline{r}_1 \times \underline{v}_1) + m_2(\underline{r}_2 \times \underline{v}_2) = \underline{h} \qquad \text{(3 integrals, total angular momentum)} \qquad (12)$$

Now multiply (9) by $\cdot\dot{\underline{r}}_1$, (10) by $\cdot\dot{\underline{r}}_2$ and add.

$$m_1\dot{\underline{r}}_1 \cdot \ddot{\underline{r}}_1 + m_2\dot{\underline{r}}_2 \cdot \ddot{\underline{r}}_2 = \frac{d}{dt} \frac{m_1 m_2}{r_{12}}$$

Integrating, we have

$$\frac{1}{2}(m_1 v_1^2 + m_2 v_2^2) = \frac{m_1 m_2}{r_{12}} + E. \text{ (1 integral, conservation of energy)} \qquad (13)$$

Subtracting (9) from (10) to get the equation of motion of the second particle with respect to the first, we have

$$\ddot{\underline{r}}_2 - \ddot{\underline{r}}_1 = - \frac{m_1 + m_2}{r_{12}^3} (\underline{r}_2 - \underline{r}_1) . \qquad (14)$$

Let $\underline{r} = \underline{r}_2 - \underline{r}_1$ and $\mu = m_1 + m_2$ so that (14) becomes

$$\ddot{\underline{r}} = - \frac{\mu}{r_{12}^3} \underline{r} \qquad \text{(Central force problem)} \qquad (15)$$

Since we now have more than 12 integrals, some must be redundant, and can be reduced to the following:

$$\underline{v}_c = \underline{\ell} \qquad (3)$$

$$\underline{r}_c = \underline{\ell} t + \underline{k} \qquad (3)$$

$$m_1(\underline{r}_1 + \underline{v}_1) + m_2(\underline{r}_2 + \underline{v}_2) = \underline{h} \qquad (3) \qquad (16)$$

$$\frac{1}{2}(m_1 v_1^2 + m_2 v_2^2) = \frac{m_1 m_2}{r_{12}} + E \qquad (1)$$

$$\frac{\underline{r}}{r} + \underline{e} = \frac{\underline{v} \times \underline{h}}{\mu} \qquad (2)$$

Returning to the problem of the time of perihelion passage, operating on both sides of the last equation in (16) by $\cdot\underline{r}$ we have

$$r + \underline{e} \cdot \underline{r} = \frac{h^2}{\mu},$$

which can be rewritten as

$$r = \frac{h^2/\mu}{1 + e \cos \omega} \qquad (17)$$

which is the polar equation of a conic section with major axis along $\underline{e}$, and ω the angle between $\underline{e}$ and $\underline{r}$.

Finally, squaring the last equation in (16) we have

$$1 + \frac{2}{r}\left\{\frac{h^2}{\mu} - r\right\} + e^2 = \frac{v^2 h^2}{\mu^2}, \quad \text{or}$$

$$\frac{v^2}{2} = \frac{\mu}{r} + \frac{(e^2 - 1)\mu^2}{2h^2} = \frac{\mu}{r} + E. \tag{18}$$

This is the conservation of energy statement, with hyperbolic motion for $e^2 > 1$, parabolic for $e^2 = 1$, elliptic for $e^2 > 1$ and $h \neq 0$.

Since $|\underline{a} \cdot \underline{b}|^2 + |\underline{a} \times \underline{b}|^2 = a^2 b^2$, substituting $\underline{r}$ and $\underline{v}$ we have

$$r^2 \dot{r}^2 + h^2 = r^2 v^2 \quad \text{or} \quad \frac{1}{2}\left(\dot{r}^2 + \frac{h^2}{r^2}\right) = \frac{\mu}{r} + E.$$

But $\frac{v^2}{2} = \frac{\mu}{r} + E$ is valid even if $h = 0$, since $\dot{\underline{r}} = \frac{-\mu}{r^2}\,\underline{r}$.

Without integrating, (18) shows

$$\frac{1}{2}\frac{h^2}{r^2} \leq \frac{\mu}{r} + E. \tag{19}$$

Multiplying (19) by r^2, where $r = |\underline{r}_2 - \underline{r}_1|$, we have

$$\frac{1}{2}h^2 \leq \mu r + Er^2 \tag{20}$$

From (20) we see that if $r \longrightarrow 0, \Longrightarrow h = 0$. In the n-body problem, all bodies cannot collide simultaneously unless the total angular momentum is 0.

Multiplying (18) by $2r^2$ and simplifying, we have

$$(r\dot{r})^2 + h^2 = 2(\mu r + Er^2) \tag{21}$$

Let $r = a(1 - e\cos\omega)$ so that $t = t_o + c(\omega - e\sin\omega)$. (22)
(22) shows that r and t can be expressed parametrically as functions of ω.

However, using an analytic approach, let

$$\dot{r} = \frac{dr}{d\mu}\,\frac{d}{dt}$$

so (21) becomes

$$\left(r\frac{d\mu}{dt}\right)^2\left(\frac{dr}{d\mu}\right)^2 + h^2 = 2(\mu r + Er^2). \tag{23}$$

Let $\left(r\frac{d\mu}{dt}\right)^2 = k^2$ and change variables so that $\frac{dr}{d\mu} \neq 0$. $\left(r' = \frac{dr}{d\mu}\right)$

Differentiating (23),

$2k^2r'r'' = 2\mu r' + 4Err'$, and

$k^2r'' = 2\mu + 4Er$, with solution of the form
$r = A + B\cos\mu$, $k = |2E|$.

To find t as a function of μ, $r\frac{d\mu}{dt} = k$, or $dt = \frac{1}{k}\,r\,d\mu$, so that

$$t = \frac{1}{k}\int r\,d\mu\ .$$

In the three body problem, either

a) none of the bodies collide. $(h \neq 0)$

b) two bodies collide. $(h \neq 0)$

c) all three collide. $(h = 0)$

If two collide, introduce appropriate time variables such that there exist solutions without singularities.

$$\mu = \int^{t} u \, dt$$

Problems:

1) If $h = 0$, find the time of collision in terms of the initial conditions for the 2 body problem.

2) Assuming $\frac{1}{r^3}$ law, show collision can occur even if $h \neq 0$.

3) Define $U = \frac{\mu}{r}$ and $E < 0$, $h \neq 0$, prove

$$\frac{1}{p} \int_{0}^{p} U(\tau) \, d\tau = -2E.$$

Lecture Two

Recalling the equation

$$(rr')^2 + h^2 = 2\mu r + 2Er^2, \quad \text{where} \tag{1}$$

$E = \frac{(e^2 - 1)\mu^2}{h^2}$, $k = r\frac{d\mu}{dt}$, then

$$k^2r'^2 + h^2 = 2\mu r + 2Er^2 . \tag{2}$$

Differentiating (2) and dividing by 2r' yields

$$k^2r'' = \mu + 2Er . \tag{3}$$

For the case where $E = 0$ in (3), arbitrarily choose $k^2 = \mu$, so that

$$k^2r'' = \mu \text{ and } r = \frac{\mu^2}{2} + \ldots , \tag{4}$$

$$t = \frac{\mu^3}{6} + \ldots , \tag{5}$$

and we obviously have parabolic motion.

For $E < 0$, choose $k^2 = 2|E|$, yielding $r'' + r = \frac{\mu}{k^2}$.

Thus $r = \frac{\mu}{k^2} + A \cos(w - B)$.

Since $k = r\frac{d\mu}{dt}$, choose $B = 0$ so we have

$$r = \frac{\mu}{k^2} + A \cos\mu \tag{6}$$

$$t = \mu + c \sin\dot{\mu} . \qquad (\text{Note: } t = \int rd\mu .) \tag{7}$$

For $E > 0$, choose $k^2 = 2E$, yielding finally

$$r = \frac{\mu}{k^2} + A \cosh \mu , \tag{8}$$

$$t = \sinh \mu . \tag{9}$$

Let us now discuss two basic problems of interest in the two body problem:

1) For those orbits in which the masses are separating as $t \longrightarrow \infty$, how large is r?

2) If $h = 0$, and for some time $t = t_1$, a collision occurs, how small is r?

Considering these problems in the order presented, from

(4) and (5),	$r \sim t^{2/3}$	for $E = 0$,
(8) and (9),	$r \sim t$	for $E > 0$,
(6) and (7),	r bounded	for $E < 0$.

Now consider the problem when $h = 0$ and collision occurs at $t = t_1$. In short, in what way is r related to $(t_1 - t)$ as $t \longrightarrow t_1$. From $\ddot{x} = \frac{-\mu}{x^2}$, multiplying by $\dot{x}$ and integrating, we have $\frac{\dot{x}^2}{2} = \frac{\mu}{x} + E$.

Multiplying now by x and taking $\lim_{x\to 0}$, we find

$$\lim_{\substack{x\to 0\\ t\to t_1}} x\ddot{x}^2 = 2\mu \tag{10}$$

Assume there exists an α such that $x \sim (t_1 - t)^{\alpha}$. Substitution in (10) yields

$$\alpha^2(t - t_1)^{\alpha}(t_1 - t)^{2\alpha - 2} \longrightarrow 2\mu, \quad \text{or}$$

$$(t_1 - t)^{3\alpha - 2} \longrightarrow 2\mu, \quad \text{or}$$

$$x \sim (t_1 - t)^{2/3} \text{ as } t \longrightarrow t_1$$

A further interesting property (Bertrand 1873) is that one has a particle in a circular orbit and the initial conditions are changed, only the inverse square law $(-\frac{\mu}{r^2})$ and the law (μr) will yield a new closed orbit. In the solar system under a μr law, the planets would move in elliptic orbits with the sun at the center, and with a common period.

Let us now discuss the n-body problem under an arbitrary law f(r). The equations of motion become

$$m_k \ddot{\underline{r}}_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} m_j m_k f(r_{jk}) \frac{\underline{r}_j - \underline{r}_k}{\underline{r}_{jk}} \quad , \tag{11}$$

If $f(r)$ is a real, analytic function, for each $r = r_i$, there exists a power series expansion of $f(r_i)$ in the neighborhood of r_i which satisfies (11) and the given initial conditions and is unique. Summing (11) over all values of k,

$$\sum_{k=1}^{n} m_k \ddot{\underline{r}}_k = 0 \tag{12}$$

Again using the concept of the mass center, with

$$\underline{r}_c = \frac{1}{M}(m_1 \underline{r}_1 + m_2 \underline{r}_2 + \ldots) \ , \ \ddot{\underline{r}}_c = 0 \ , \ \dot{\underline{r}}_c = \underline{k} \ , \ \underline{r}_c = \underline{k}t + \underline{h} \ ,$$

note $\underline{r}_k = \underline{r}_k - \underline{r}_c$ but (11) is unchanged, so we have

$$\sum_{k=1}^{n} m_k \underline{r}_k = 0 \ , \quad \sum_{k=1}^{n} m_k \underline{v}_k = 0.$$

The order of the system has now been reduced to $6n - 6$.

If $f(r) = r$, (11) can be reduced to

$$\ddot{\underline{r}}_k = - M \underline{r}_k \ , \quad k = 1, \ldots, n. \tag{13}$$

(13) implies all the masses satisfy the same differential equation, but fails to recognize that perhaps two of the masses may collide. From (13), the motion is elliptic or linear, and

$$\underline{r}_k = \underline{A}_k \cos \omega t + \underline{B}_k \sin \omega t, \quad \omega = \sqrt{M} \tag{14}$$

For the situation $f(r) = \frac{1}{r^2}$ where $n = 2$ or 3, if the solution to (11) ceases to be analytic at some time $t = t_1$, a collision has occurred. For $n > 3$ the problem remains unsolved.

However, Painlevé has shown that if in some finite time $t = t_1$, a singularity occurs, then

$$\min_{t \to t_1} \underline{r}_{jk} = 0,$$

where we have $\frac{n(n-1)}{2}$ distances $\underline{r}_{jk}$.

Returning to (11), crossing by $\underline{r}_k$ and integrating we finally get

$$\sum_{k=1}^{n} m_k (\underline{r}_k \times \underline{v}_k) = \underline{h} \ , \tag{15}$$

subject to $\sum_{k=1}^{n} m_k \underline{r}_k = 0$ and $\sum_{k=1}^{n} m_k \underline{v}_k = 0$. The system has now been reduced to $6n - 9$.

For the final reduction, define $\mu(r)$ such that $\mu'(r) = -f(r)$. Form the self potential

$$U = \sum_{1 \le j < k \le n} m_j m_k \mu(r_{jk}). \tag{16}$$

Relating (16) to (11), we have

$$m_k \ddot{\underline{r}}_k = \text{grad}_k U \quad . \tag{17}$$

Multiplying (17) by $\dot{\underline{r}}_k \cdot$ and integrating,

$$\tfrac{1}{2} \sum_{k=1}^{n} m_k v_k^2 = E + U \quad . \qquad \text{(Conservation of energy).} \tag{18}$$

To specify the constant, and reduce the system to $6n - 10$, for

$$f(r) = \frac{1}{r^2} \text{ , let } \mu(r) = \int_r^{\infty} f(r)\, dr \text{ ,}$$

$$f(r) = r \text{ , let } \mu(r) = -\int_0^{r} f(r)\, dr \text{ ,}$$

$$f(r) = \frac{1}{r} \text{ , let } \mu(r) = -\int_1^{r} f(r)\, dr \text{ .}$$

To obtain the Lagrange - Jacobi form of $\ddot{I}$, recall that

$$\dot{I} = \sum_{k=1}^{n} m_k \underline{r}_k \cdot \underline{v}_k$$

and differentiate.

$$\ddot{I} = \sum_{k=1}^{n} m_k (\underline{r}_k \cdot \ddot{\underline{r}}_k + \underline{v}_k^2)$$

$$= 2T + \sum_{k=1}^{n} m_k \underline{r}_k \cdot \ddot{\underline{r}}_k \quad . \tag{1}$$

But $\sum_{k=1}^{n} m_k \underline{r}_k \cdot \ddot{\underline{r}}_k = \sum_{k=1}^{n} \underline{r}_k \cdot \text{grad}_k U$, so we have

$$\ddot{I} = 2T + \sum_{k=1}^{n} \underline{r}_k \cdot \text{grad}_k U \quad . \tag{2}$$

A function $f(x_1,\ldots,x_m)$ is homogeneous of order k if there exists $0<\lambda$ such that

$$f(\lambda x_1,\ldots,\lambda x_m) = \lambda^k f(x_1,\ldots,x_m) \quad . \tag{3}$$

Differentiating (3) with respect to λ , and letting λ become 1 we get

$$\sum_{s=1}^{n} x_s \frac{\partial f}{\partial x_s} = kf \tag{4}$$

but this is precisely the expression for the coordinates in (2) if U is homogeneous of order ℓ. i.e.

$$\sum_{k=1}^{n} \left(X_k \frac{\partial U}{\partial X_k} + Y_k \frac{\partial U}{\partial Y_k} + Z_k \frac{\partial U}{\partial Z_k} \right) = \ell U. \qquad \text{(Virial function)} \qquad (5)$$

Thus

$$\ddot{I} = 2T + \ell U. \qquad (6)$$

Consider the effect of letting $f(r) = \frac{1}{r^p}$, $\quad -\infty < p < \infty$.

From the previous lecture, $\mu(r)$ was defined as follows:

$\mu(r) = \int_r^\infty f(r)\, dr$, if it makes sense, which in this case depends on the value of p. Allowing for various values of p,

$$\text{If } p > 1, \quad \mu(r) = \int_r^\infty \frac{dr}{r^p} = \frac{1}{p-1} \cdot \frac{1}{r^{p-1}} \quad .$$

$$\text{If } p < 1, \quad \mu(r) = -\int_0^r \frac{dr}{r^p} = \frac{1}{p-1} \cdot \frac{1}{r^{p-1}} \quad .$$

$$\text{If } p = 1, \quad \mu(r) = -\int_1^r \frac{dr}{r} = \log \frac{1}{r} \quad .$$

Thus, if $p < 1$ or $p > 1$, $\mu(r)$ and consequently U(r) is homogeneous of degree $(1 - p)$. (Recall $U = \sum_{1 \le j < k \le n} m_j m_k \mu(r_{jk})$).

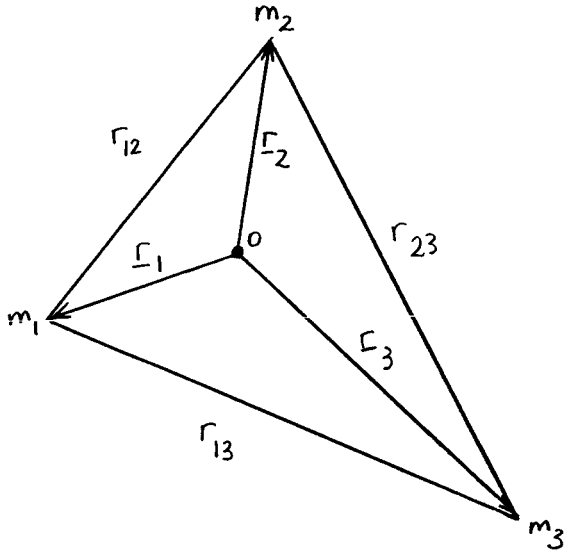
Applying the homogeniety property to (6),

$$\ddot{I} = 2T + (1 - p)U \quad \text{for} \quad f(r) = \frac{1}{r^p} \quad , \; p \neq 1. \qquad (7)$$

Since $T = U + E$ and $U = T - E$, (7) becomes

$$\ddot{I} = 2U + 2E + (1 - p)U = (3 - p)U + 2E \equiv (3 - p)T + (p - 1)E. \quad (8)$$

We are now in a position to qualitatively discuss the relationship between I, U, T and the general geometry of the problem.



Let O be the mass center of this three body system, and define

$$R(t) = \max. r_{jk}(t)$$

$$\varsigma(t) = \max. r_k(t)$$

$$r(t) = \min. r_{jk}(t),$$

where these functions by their very definition are not necessarily analytic. The use of inequalities will enable us to relate I to these new functions.

$$2I = \sum_{k=1}^{n} m_k r_k^2 .$$

By definition, each $r_k^2 \leq \varsigma^2(t)$

$$\therefore 2I \leq \sum_{k=1}^{n} m_k \varsigma^2 \quad , \text{ or}$$

$$I \leq \frac{M}{2} \varsigma^2(t) \quad . \quad (9)$$

Similarly,

$\sum m_k r_k^2 \geq m \sum r_k^2 \geq m \varsigma^2$ where $m = \min m_k$. Combining this result with (9),

$$\frac{m}{2}\zeta^2 \leq I \leq \frac{M}{2}\zeta^2 \qquad (10)$$

Inequality (10) tells us I and ζ^2 are of the same order.

Now how is I related to R? An alternate form for I, valid for the case where the center of mass is fixed, and useful in this treatment, is

$$\sum_{k=1}^{n} m_k(\underline{r}_k - \underline{r}_j)^2 = \sum_{k=1}^{n} m_k r_k^2 + r_j^2 \sum_{k=1}^{n} m_k - 2\sum_{k=1}^{n} m_k \underline{r}_k \cdot \underline{r}_j \quad . \quad (11)$$

But with a fixed center of mass, the last term in (11) vanishes, and

$$\sum_{k=1}^{n} m_k (\underline{r}_k - \underline{r}_j)^2 = 2I + Mr_j^2 \qquad (12)$$

Multiplying (12) by m_j and summing with respect to j, we finally get

$$\frac{1}{2M} \sum_{1 \leq j < k \leq n} m_{jk} r_{jk}^2 = I \qquad (m_{jk} = m_j \cdot m_k)$$

But from the definition of R(t),

$$I \leq \frac{1}{2M} R^2(t) \sum m_{jk} \quad \text{or simply}$$

$$I \leq AR^2(t) \quad \text{where A is a constant.} \qquad (13)$$

Applying similar techniques as in (10) we find there are constants A and B such that

$$BR^2(t) \leq I \leq AR^2(t) \tag{14}$$

so I is of the order of $R^2(t)$. A general conclusion is that if for $t \longrightarrow a$ one of the quantities I, R or $\zeta \longrightarrow \infty$, they all $\longrightarrow \infty$.

Let us now show the relation between r, U and T for $1 < p < 3$. Since $r_{jk} \geq r$, $\frac{1}{r_{jk}} \leq \frac{1}{r}$, and from the defintion of U,

$$\frac{B}{r^{p-1}} \leq U = \sum_{i \leq j < k \leq n} \frac{m_{jk}}{(p-1)\, r_{jk}^{p-1}} \leq \frac{A}{r^{p-1}} \tag{15}$$

A conclusion for (15) and the preceding work is that if $t \longrightarrow a$ one of the quantities $\frac{1}{r}$, U, T or $\ddot{I}$ approaches ∞, they all do.

It is impossible for all the bodies to collide simultaneously after an infinite time. To prove this statement, assume the contrary, i.e., all $r_k \longrightarrow 0$ implies $R(t) \longrightarrow 0$ for $t \longrightarrow \infty$.

$r \longrightarrow 0$ implies $\frac{1}{r} \longrightarrow \infty$, which implies $\ddot{I} \longrightarrow \infty$. If $\ddot{I} \longrightarrow \infty$, at some time $\ddot{I} > 0$. For simplicity let

$\ddot{I} > A$ for some t and with $A > 0$.

$\ddot{I} > A$, integrated twice, yields

$$I > \frac{A_1 t^2}{2} + C_1 t + C_2 \tag{16}$$

But (16) tells us that $I \longrightarrow \infty$ which implies $R \longrightarrow \infty$ from (13). But this is contrary to our hypothesis.

Consider the problem of less than n bodies colliding after some time $t = a$. Let $n = 3$, $f(r) = \frac{1}{r^2}$. Then Chazy proved (1923) that it is impossible for a particular pair of masses to collide as $t \longrightarrow \infty$ if there exists a quantity $\delta > 0$ such that both remaining distances are always greater than or equal to δ. i.e., $r_{12} \not\longrightarrow 0$ as $t \longrightarrow \infty$ if there exists δ such that r_{23} and $r_{31} \geq \delta > 0$.

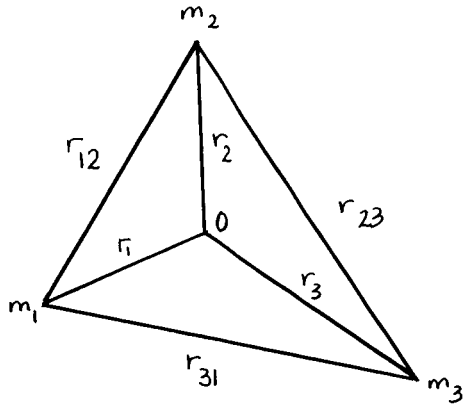


Figure 1

Pollard Theorem:

Recalling the definitions:

$$R(t) = \text{max. } r_{jk}(t)$$
$$\varsigma(t) = \text{max. } r_{k}(t)$$
$$r(t) = \text{min. } r_{jk}(t).$$

Then for $n = 0$, $r(t) \not\longrightarrow 0$ as $t \longrightarrow \infty$.

Proof: Assume $r \longrightarrow 0$. This implies there exists an $r_{jk} \longrightarrow 0$ for some particular j and k. If no particular pair becomes and <u>remains</u> the minimum pair, this implies at least two r_{jk} are alternately the minimum. Let them be r_{12} and r_{23}. Then when they exchange positions, i.e., $r_{12} < r_{23} \longrightarrow r_{23} < r_{12}$, there exists a time t_m such

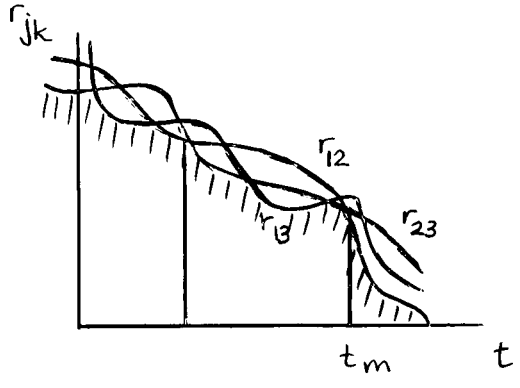


Figure 2

that $r_{12} = r_{23}$. But each time they exchange position, these exists sequence of times t_n for which $r_{12} = r_{23}$. But from Figure 1: $r_{31} \leq r_{12} + r_{23}$. Thus $r_{31}(t_n) \longrightarrow 0$. Remember the form $I = \frac{1}{2M}(m_{12}r_{12}^2 + m_{23}r_{23}^2 + m_{31}r_{31}^2)$. (1) This implies $I(t_n) \longrightarrow 0$ as $r_{jk} \longrightarrow 0$. But for $r(t) > 0$, we have shown $I > At^2$. Following the same logic of yesterday's lecture, we arrive at a contradiction. Thus if $r(t) \longrightarrow 0$, a fixed r_{jk} will eventually become and remain the $r(t)$ of our definition.

From (1) and previous results, assuming it is r_{12} that becomes the min. so $r_{12} \longrightarrow 0$, we have

$$\frac{1}{2M}(m_{23}r_{23}^2 + m_{31}r_{31}^2) \geq At^2. \quad \text{Assume } m_{23} > m_{31}.$$

Then

$$m_{23}(r_{23}^2 + r_{31}^2) \geq At^2 \quad \text{or}$$
$$r_{23}^2 + r_{31}^2 > Bt^2 \ . \tag{2}$$

Now let us show both r_{23} and r_{31} are greater than some multiple of t^2.

$$\left| r_{23} - r_{31} \right| \leq r_{12} \Longrightarrow r_{23} - r_{31} \longrightarrow 0 \Longrightarrow r_{23} = r_{31} + 2\epsilon \ ,$$

where $\epsilon \longrightarrow 0$ as $t \longrightarrow \infty$. Rewriting (2) in terms of r_{23}, we have

$$(r_{31} + 2\epsilon)^2 + r_{31}^2 > Bt^2 \ . \tag{3}$$

Since (3) depends on r_{31}, i.e., $\epsilon^2 \longrightarrow 0$, (3) can be represented as

$$r_{31}^2 + 2\epsilon\, r_{31} + \epsilon^2 > Bt^2, \quad \text{or}$$

$$(r_{31} + \epsilon)^2 > Bt^2 \Longrightarrow r_{31} + \epsilon > Ct^2 \text{ and as } \epsilon \longrightarrow 0$$

$$r_{31} > Dt.$$

The same argument applies to r_{23} as $r_{12} \longrightarrow 0$. Thus both r_{31} and r_{23} increase more rapidly than t. From

$$m_3 \ddot{\underline{r}}_3 = \frac{m_1 m_3}{r_{13}^3}(\underline{r}_1 - \underline{r}_3) + \frac{m_2 m_3}{r_{23}^3}(\underline{r}_2 - \underline{r}_3) \ ,$$

$$\left|\ddot{\underline{r}}_3\right| \leq \frac{m_1}{r_{13}^2} + \frac{m_2}{r_{23}^2} \leq \frac{A}{t^2} \ . \tag{4}$$

Integrate (4) between $t = t_1$ and $t = t_2$, $t_1 < t_2$.

$$\left| \int_{t_1}^{t_2} \ddot{\underline{r}}_3 \, dt \right| \leq \int_{t_1}^{t_2} \left| \ddot{\underline{r}}_3 \right| dt \leq C \left(\frac{1}{t_1} - \frac{1}{t_2} \right) \ , \text{ or}$$

$$\left| \underline{v}_3(t_2) - \underline{v}_3(t_1) \right| \leq C \left(\frac{1}{t_1} - \frac{1}{t_2} \right) \longrightarrow 0 \text{ as } t_1 \text{ and } t_2 \longrightarrow \infty \tag{5}$$

However, Cauchy proved that if there exists an $f(t)$, $t > 0$ such that $f(t_1) - f(t_2) \longrightarrow 0$ as t_1 and $t_2 \longrightarrow \infty$, then $\lim_{t \to \infty} f(t)$ exists. Applying this to (5), $\lim_{t \to \infty} v_3 (\infty)$ exists.

Now let $t_2 \longrightarrow \infty$, so (5) becomes

$$\left| \underline{v}_3 (\infty) - \underline{v}_3 (t) \right| \leq C \cdot \frac{1}{t} \ . \tag{6}$$

Dropping the subscript in (6) and integrating, with $t_1 = 1$ (arbitrary choice)

$$\left| \int_1^t \left[\underline{v}_3(t)\, dt - \underline{v}_3(\infty) \right] dt \right| \leq \int_1^t \left| \underline{v}_3(t) - \underline{v}_3(\infty) \right| dt \leq$$

C log t; becomes

$$\left| \underline{r}_3(t) - \underline{v}_3(\infty)t + C_1 \right| \leq C \log t. \tag{7}$$

Divide (7) by t and let $t \longrightarrow \infty$. Since $\frac{\log t}{t}$ and $\frac{c}{t} \longrightarrow 0$,

$$\lim_{t \to \infty} \frac{\underline{r}_3(t)}{t} \longrightarrow \underline{v}_3(\infty). \tag{8}$$

Since $\sum_i m_i \underline{r}_i = 0$, $\sum_i \frac{m_i r_i}{t} = 0$, the last term of

$$\frac{m_i \underline{r}_i + m_2 \underline{r}_2}{t} + \frac{m_3 \underline{r}_3}{t} = 0 \text{ vanishes.} \tag{9}$$

Thus $\lim_{t \to \infty} \frac{m_1 \underline{r}_1 + m_2 \underline{r}_2}{t}$ exists.

Assuming $\underline{r}_2 - \underline{r}_1 \longrightarrow 0$ and dividing by $\frac{t}{m_1}$,

$$m_1 \frac{\underline{r}_2}{t} - m_1 \frac{\underline{r}_1}{t} \longrightarrow 0. \tag{10}$$

Combining (9) and (10) we now have

$(m_1 + m_2) \frac{\underline{r}_2}{t}$, $\frac{\underline{r}_1}{t}$ and $\frac{\underline{r}_2}{t}$ all having limits.

Since

$$\frac{I}{t^2} = \frac{m_1 r_1^2}{t^2} + \frac{m_2 r_2^2}{t^2} + \frac{m_3 r_3^2}{t^2}, \text{ we now have}$$

$$\lim_{t \to \infty} \frac{I}{t^2} \text{ exists and is finite.} \tag{11}$$

From previous results, for some $t > t_0$ and $A > 0$,

$$I > \frac{At^2}{2} + Bt + C. \tag{12}$$

Dividing by t^2 and letting $t \longrightarrow \infty$, (12) becomes

$$\lim_{t \to \infty} \frac{I}{t^2} > \frac{A}{2}.$$

Since A is arbitrary, let $A \longrightarrow \infty$; so that

$$\lim_{t \to \infty} \frac{I}{t^2} = \infty.$$

This implies I increases more rapidly than the quadratic t^2, and contradicts (11). Therefore $r \not\longrightarrow 0$ as $t \longrightarrow \infty$.

In the n-body problem, a simultaneous collision of all n bodies implies that the total angular momentum is zero. ($\underline{h} = 0$).

Proof: Such a collision implies $R(t) \to 0$. We have previously shown $R \not\to 0$ after an infinite time, so there exists a time $t = t_1 < \infty$ at which the collision must occur; but this is impossible unless $\underline{h} = 0$. To prove this last remark, recall

$$\underline{h} = \sum_{k=1}^{n} m_k (\underline{r}_k \times \underline{v}_k), \text{ so that}$$

$$\underline{h} \leq \sum_k m_k r_k v_k \quad . \tag{13}$$

Since it is true that

$$\left| \sum ab \right|^2 \leq \sum a^2 . \quad \sum b^2,$$

(13) can be written as

$$\left| \underline{h} \right| \leq \sum_k (\sqrt{m_k}\ r_k) \quad \sum_k (\sqrt{m_k}\ v_k), \quad \text{or}$$

$$\left| \underline{h} \right|^2 \leq \sum_k m_k r_k^2 \quad \sum_k m_k v_k^2 = 4\, I\, T. \tag{14}$$

Using $T = \ddot{I} - E$, (14) becomes

$$\left| \underline{h} \right|^2 \leq 4I(\ddot{I} - E) \tag{15}$$

But at some time t_1, $I \longrightarrow 0$, and $R \longrightarrow 0 \Longrightarrow r \longrightarrow 0$. Similarly $\ddot{I} \longrightarrow 0$ implies that at some time $\ddot{I} > A > 0$. So the plot of I vs. t must be concave upwards. But this means $\dot{I} < 0$, or $-\dot{I} > 0$. Multiplying (15) by $-\dot{I}/I$,

$$\frac{\left| \underline{h} \right|^2}{I} (-\dot{I}) \leq -4\dot{I}(\ddot{I} - E). \tag{16}$$

Integrating (16) in the neighborhood of t_1,

$$h^2 \log \frac{1}{I} \leq 4EI - 2\dot{I}^2 + K. \tag{17}$$

But $2\dot{I}^2$ is negligible, $\frac{1}{I} \longrightarrow \infty$, $\log \frac{1}{I} \longrightarrow \infty$ so is eventually > 0. Dividing (17) by $\log \frac{1}{I}$,

$$h^2 \leqq \frac{4EI + K}{\log \frac{1}{I}} \qquad \text{for } t \sim t_1 .$$

Now as $t \longrightarrow t_1$, the denominator $\longrightarrow \infty$, $I \longrightarrow 0$ and the numerator $\longrightarrow K$.

Thus $h^2 \longrightarrow 0$ as $t \longrightarrow t_1$, or

the total angular momentum vanishes if these exists a simultaneous collision of all masses.

Tauberian Theorem and Condition

Consider the problem of a given function $f(x)$, $x > 0$ such that $f(x) \sim Ax^2$, i.e., $\lim_{x\to\infty} \frac{f(x)}{x^2} = A$, and let us question if it is true that $f'(x) \sim 2Ax$.

The converse statement

$f'(x) \sim 2Ax \Longrightarrow f(x) \sim Ax^2$ is true, but it is not necessarily true that given an asymptotic function, one can differentiate with the result an asymptotic function. It is this irreversibility that led to the concept of the Tauberian condition, which is that additional information required to obtain reversibility in the above limits.

$$\lim_{x\to\infty} \frac{f'(x)}{2x} = A \Longrightarrow \text{there exists } \epsilon > 0 \text{ and an } x_0 \text{ such that}$$

$$\left| \frac{f'(x)}{2x} - A \right| < \epsilon \quad \text{as } x \longrightarrow x_0. \tag{1}$$

Multiply (1) by $2x$, integrate with respect to x, divide by x^2 and let $x \longrightarrow \infty$.

$$\overline{\lim_{x\to\infty}} \left| \frac{f(x)}{x^2} - A \right| \leq \frac{\epsilon}{2}. \tag{2}$$

Since ϵ is arbitrary, let $\epsilon \longrightarrow 0$ so (2) becomes

$$\lim_{x\to\infty} \frac{f(x)}{x^2} = A \quad \text{or} \quad f(x) \sim Ax^2. \tag{3}$$

Now, to show $f(x) \sim Ax^2 \Longrightarrow f'(x) \sim 2Ax$ we must introduce a Tauberian condition.

Landau Theorem (1906)

Case 1. $A = 0$. If $f(x) \sim Ax^2$, and $f''(x) \geq C \quad -\infty$, then $f'(x) \sim 2AX$. Consider $f(x + \epsilon x)$, $\epsilon \gtrless 0$, and its expansion in Taylor series with a remainder, up to second order terms.

$$f(x + \epsilon x) = f(x) + \epsilon x f'(x) + \frac{\epsilon^2 x^2}{2} f''(\xi), \quad x < \xi < x + \epsilon x. \quad (4)$$

Using Landau's Theorem in (4),

$$f(x + \epsilon x) \geq f(x) + \epsilon x f'(x) + \frac{\epsilon^2 x^2}{2} \cdot C. \quad (5)$$

Dividing (5) by x^2 and taking lim as $x \longrightarrow \infty$, we have left

$$0 \geq \overline{\lim_{x \to \infty}} \frac{\epsilon f'(x)}{x} + \frac{\epsilon^2}{2} C. \quad (6)$$

In (6), for $\epsilon > 0$, divide by ϵ and let $\epsilon \longrightarrow 0$ so (6) becomes

$$\overline{\lim_{x \to \infty}} \frac{f'(x)}{x} \leq 0. \quad (7)$$

In (6), for $\epsilon < 0$, divide by ϵ, (reverse inequality) and (6) becomes

$$\underline{\lim_{x \to \infty}} \frac{f'(x)}{x} \geq 0. \quad (8)$$

But (7) and (8) imply

$$\lim_{x \to \infty} \frac{f'(x)}{2x} = 0$$

The same conclusion can be reached for $f''(x) \leqslant C < \infty$ by using -f for f in the above argument.

Case 2. $A \neq 0$. Define g(x) such that $g(x) = f(x) - Ax^2$, so that

$$\frac{g(x)}{x^2} = \frac{f(x)}{x^2} - A. \tag{9}$$

The hypothesis $f(x) \sim Ax^2 \Longrightarrow \frac{g(x)}{x^2} \longrightarrow 0.$

From (9), $\frac{g''(x)}{x^2} = f''(x) - 2A$, where $f''(x) \geqslant C > -\infty$.

But the argument for the case A = 0 now applies to g''(x), so

$$\lim_{x \to \infty} \frac{g'(x)}{2x} = 0. \tag{10}$$

By definition, $g'(x) = f'(x) - 2AX$, so that

$$\frac{g'(x)}{2x} = f'(x) - A. \tag{11}$$

Using (10) and (11), $f'(x) \sim 2AX$

von Clausius' Theorem $\left[f(r) = \frac{1}{r^2}\right]$

If a system is bounded in size and velocity, then both the kinetic energy and potential energy have limits in the average sense. i.e.

$$\hat{U} = \lim_{t \to \infty} \int_0^t U \, dt = -2E, \tag{12}$$

$$\hat{T} = \lim_{t\to\infty} \int_0^t T\,dt = -E. \qquad (13)$$

(Classical Virial).

Proof: If (12) is true, (13) follows, since

$T = \hat{U} + E$, and since E is constant is equal to its average value. Thus $\hat{T} = \hat{U} + E$. In words, $\hat{T}$ is redundant in view of the conservation of energy.

Now to establish (12). Begin with the Lagrange - Jacobi identity $\ddot{I} = U + 2E$, integrate once and divide by t.

$$\frac{\dot{I}}{t} = \frac{1}{t}\int_0^t U\,dt + 2E + \frac{C}{t}.$$

Remembering $\dot{I} = \sum_{k=1}^{n} m_k (\underline{r}_k \cdot \underline{v}_k)$, from the hypothesis of bounded $\underline{r}_k$ and $\underline{v}_k$, $\dot{I}$ is bounded in time. Thus, as $t \longrightarrow \infty$, $\frac{\dot{I}}{t} \longrightarrow 0$, $\frac{C}{t} \longrightarrow 0$, and

$$\hat{U} = \frac{1}{t}\int_0^t U\,dt = -2E.$$

However, the fact that is some cases, i.e., parabolic case of the two body problem, $\hat{U} = -2E\ (= 0)$ even for an unbounded system, $(r \sim t^{2/3})$, Pollard has developed a stronger theorem.

Pollard Theorem $\left[f(r) = \frac{1}{r^2}\right]$

A necessary and sufficient condition that U exist and equal - 2E is that $\overline{\lim_{t\to\infty}} \frac{R(t)}{t} = 0$. In other words, if the system is bounded (note:

velocity not involved), $\frac{R(t)}{t} \longrightarrow 0$ and $\hat{U} = -2E$.

First let us show $\frac{\dot{I}}{t} \longrightarrow 0 \equiv \frac{I}{t^2} \longrightarrow 0$. We have already shown

$$\frac{\dot{I}}{t} = \frac{1}{t}\int_0^t U\,dt + 2E + \frac{C}{t} \Longrightarrow \hat{U} = -2E \text{ if and only if } \frac{I}{t} \longrightarrow 0.$$

$$\text{If } \lim_{x\to\infty} \frac{f'(x)}{2x} = 0, \text{ then } \lim_{x\to\infty} \frac{f(x)}{x^2}.$$

Applying this to functions $\dot{I}$ and I,

$$\lim_{t\to\infty} \frac{I}{t} = 0 \Longrightarrow \lim_{t\to\infty} \frac{I}{t^2} = 0 \qquad (14)$$

To prove the reverse of (14), we already know that if

$$\lim_{x\to\infty} \frac{f(x)}{x^2} = 0 \text{ and } f''(x) \geq C > -\infty, \text{ then } \lim_{x\to\infty} \frac{f'(x)}{x} = 0.$$

Thus we need to show only that $\ddot{I} \geq C > -\infty$.

$\ddot{I} = U + 2E \geq 2E > -\infty$. Thus $\ddot{I}$ may be integrated to get $\lim_{t\to\infty} \frac{\dot{I}}{t} \longrightarrow 0$. Thus $\hat{U} = -2E$ if and only if $\frac{\dot{I}}{t} \longrightarrow 0$ or $\frac{\dot{I}}{t^2} \longrightarrow 0$.

But earlier we established that

$$\frac{BR^2}{t^2} \leq \frac{I}{t^2} \leq A\frac{R^2}{t^2}.$$

From this, if $\frac{I}{t^2} \longrightarrow 0$, $\frac{R^2}{t^2}$ and $\frac{R}{t} \longrightarrow 0$. Thus

$\hat{U} = -2E$ if and only if $\lim_{t\to\infty} \frac{I}{t} = 0$, or

$$\lim_{t\to\infty} \frac{I}{t^2} = 0, \text{ or}$$

$$\lim_{t\to\infty} \frac{R(t)}{t} = 0.$$

Theorem: If $\hat{T}$ exists and equals 0, then $E = 0$.

Proof: $T = U + E$, so if $\hat{T}$ exists, so does $\hat{U}$ and $\hat{T} = 0$ by hypothesis. Thus

$$\hat{T} = \hat{U} + E, \text{ or } \hat{U} = -E.$$

But $\hat{U} > 0$, so $-E \geqq 0$ or $E \leqq 0$. (15)

From $\ddot{I} = T + E$, integration once and division by t gives

$$\frac{\dot{I}}{t} = \frac{1}{t}\int_0^t T\,dt + E + \frac{C}{t} . \tag{16}$$

As $t \longrightarrow \infty$, if $T = 0$, the integral in (16) must vanish, as does $\frac{C}{t}$.

$$\text{Thus: } \lim_{t\to\infty} \frac{I}{t} = E \text{ so } \dot{I} \sim Et. \tag{17}$$

Integration of this asymptotic function gives

$$I \sim \frac{Et^2}{2}, \text{ or}$$

$$\lim_{t\to\infty} \frac{I}{t^2} = \frac{E}{2} .$$

But $I > 0$, so $E \geq 0$. Combining this result with (15) we have

$$E = 0.$$

Incidentally, the theorem is true for all $p \neq 1$. What happens if $E < 0$? Since $T = U + E$, and $T \geq 0$, by hypothesis

$$U + E \geq 0 \quad \text{or} \quad U \geq -E.$$

Thus $U \geq |E|$, but $\frac{A}{r} \geq U$, so $\frac{A}{r} \geq |E| \Longrightarrow \frac{A}{|E|} \geq r$. If $E < 0$, r is bounded but the maximum could conceivably $\longrightarrow \infty$.

Theorem: If $E > 0$, $n = 3$, a particle escapes.

Lemma 1. $4EI - \dot{I}^2 < C$ as $t \longrightarrow \infty$. From the Lagrange - Jacobi equation $\ddot{I} = 2E + U$, we conclude $\dot{I} \longrightarrow \infty$ as $t \longrightarrow \infty$. Thus, for some time $\dot{I}$ becomes positive, and define this as $\dot{I}(0)$. By integrating $\ddot{I}$,

$$\dot{I} = \int_0^t U \, d\tau + 2Et + \dot{I}\,\delta, \quad \text{or}$$

$$I = \int_0^t (t - \tau) \, d\tau \, U(\tau) + Et^2 + kt + \mathfrak{l} \tag{1}$$

Multiplying (1) by $4E$, we can write

$$4EI < 4Et \int_0^t U(\tau) \, d\tau + 4E^2t^2 + 4Ekt + C, \text{ and}$$

$$\dot{I}^2 > 4E^2t^2 + 4Et \int_0^t U(\tau) \, d\tau + 4Ekt. \tag{2}$$

Reversing the signs (and sense) of (2) we have

$$4EI - \dot{I}^2 < C.$$

Define $J = \sum_{i \le j < k \le n} m_{jk} r_{jk}$, where we now know

$$BR_{jk} < J < AR_{jk}$$

Lemma 2. $\lim_{t \to \infty} \frac{1}{UJ} = \mathfrak{l}$ exists.

$$UJ = \sum \frac{m_j k}{r_j k} \sum m_{pq} r_{pq}, \qquad \begin{matrix} 1 \leq j < k \leq n, \\ 1 \leq p < q \leq n. \end{matrix}$$

$$= \sum \sum m_{jk} m_{pq} \frac{r_{pq}}{r_{jk}} \quad . \tag{3}$$

Differentiating (3) with respect to time,

$$(UJ)' = \sum \sum m_{jk} m_{pq} \frac{r_{jk}\dot{r}_{pq} - r_{pq}\dot{r}_{jk}}{r_{jk}^2} \quad . \tag{4}$$

But $r_{jk} \geq r$ by definition, so $\frac{1}{r_{jk}} \leq \frac{1}{r}$. From (4), with $\frac{C}{r} \leq U$

$$\left|(UJ)'\right| \leq CU^2 \sum \sum \sqrt{m_{jk} m_{pq}} \sqrt{m_{jk} m_{pq}} \left| r_{jk}\dot{r}_{pq} - r_{pq}\dot{r}_{jk} \right| \quad . \tag{5}$$

Square (5), use Landau's inequality property and note that in the expansion of $\left| \qquad \right|^2$ we get a middle term

$$-2 \sum m_{jk} r_{jk} \dot{r}_{jk} \sum m_{pq} r_{pq} \dot{r}_{pq} = -4\dot{I}^2.$$

Thus (5) becomes

$$\left|(UJ)'\right|^2 \leq CU^4 (4J^2 U + C_1) \qquad\qquad (J^2 > I^2)$$

Division by $U^5 J^5$,

$$\frac{\left|(UJ)'\right|^2}{U^5 J^5} \leq C \cdot \frac{1}{J^3} + \frac{C_1}{J^4} \cdot \frac{1}{UJ} \tag{6}$$

But $\frac{1}{U} \leq Cr$, so $\frac{1}{J} \leq \frac{C}{R}$. Thus $\frac{1}{UJ} \leq \frac{Cr}{R} \leq C$. This implies $\frac{1}{UJ}$ is bounded, and will be dropped from (6). Similarly, $I > At^2 \Longrightarrow J^2 > At^2 \Longrightarrow$ $J > At$. Finally (6) can be reduced to

$$\frac{\left|(UJ)'\right|}{(UJ)^{5/2}} \leqslant \frac{B}{t^{3/2}} \tag{7}$$

Integration of (7) gives

$$\left| \frac{1}{(UJ)_1^{3/2}} - \frac{1}{(UJ)_2^{3/2}} \right| \leqslant B \left| \frac{1}{t_1^{1/2}} - \frac{1}{t_2^{1/2}} \right| . \tag{8}$$

But as t_1 and $t_2 \longrightarrow \infty$ independently, the right side of (8) is bounded, and

$$\frac{1}{(UJ)^{3/2}} \longrightarrow \lim \text{ as } t \longrightarrow \infty \text{ or } \lim_{t\to\infty} \frac{1}{UJ} = \ell .$$

Assume $\ell > 0$. Then, there exists a $\delta > 0$ such that

$$\frac{1}{UJ} \geqslant \delta \geqslant 0 \text{ as } t \longrightarrow \infty .$$

$$\frac{r}{R} \geqslant \frac{C}{UJ} \geqslant \delta > 0, \text{ so } r \geqslant \delta R \geqslant Ct. \tag{9}$$

(9) shows that as $t \longrightarrow \infty$, $r \longrightarrow \infty$ so all the particles escape.

Assume $\ell = 0$. Then at some time, some r_{jk} (say r_{12}) becomes the minimum, i.e., $r_{12} = r$. If another r_{jk} swaps with r_{12}, each time a new r_{jk} becomes the minimum, $r_{12} = r_{23}$. (Assuming r_{23} is other minimum). Then $r_{31} = R$. But

$$\left| r_{12} - r_{31} \right| \leqslant r_{23}, \text{ or}$$
$$\left| r - R \right| \leqslant r. \tag{10}$$

Division of (10) by R,

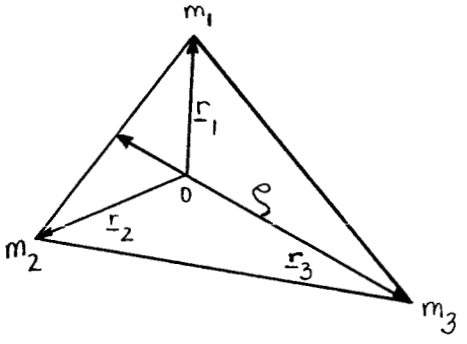
$\left|\frac{r}{R} - 1\right| \leq \frac{r}{R}$. But $\lim_{t\to\infty} \frac{r}{R} = 0$, which implies $\left| - 1 \right| \leq 0$. Therefore there exists a min r_{jk}; call it r_{12}.

Introduce Jacobi coordinates ς and r, where

$$\underline{\varsigma} = \underline{r}_3 - \frac{m_1\underline{r}_1 + m_2\underline{r}_2}{m_1 + m_2}$$

$$\underline{r} = \underline{r}_2 - \underline{r}$$

$$m_1\underline{r}_1 + m_2\underline{r}_2 + m_3\underline{r}_3 = 0$$



The above set of equations may be solved to get $\underline{r}_1$, $\underline{r}_2$ and $\underline{r}_3$ as linear functions of $\underline{r}$. Such a manupulation would show

$$I = A\varsigma^2 + Br^2 , \qquad (11)$$

where A and B are functions only of the masses.

Let $r_{23} = R$, $r_{12} = r$. Then $\left| R - \varsigma \right| \leq r$, or

$$\left|1 - \frac{\varsigma}{R}\right| \leq \frac{r}{R} .$$

But if $\Omega = 0$, $\frac{r}{R} \leq \frac{1}{UJ} \Longrightarrow \frac{\varsigma}{R} \longrightarrow 1$, $\varsigma \sim R$.

Since $R \geq Ct$, $\varsigma \sim Ct$ so ς becomes unbounded as $t \longrightarrow \infty$. But rewriting the form for ς ,

$$\zeta = \frac{M\underline{r}_3 - \sum m_i \underline{r}_i}{M - m_3} = \frac{M}{M - m_3}\,\underline{r}_3 = c\underline{r}_3 \,. \qquad (12)$$

Since, as $t \longrightarrow \infty$, $\zeta \longrightarrow \infty$, $r_3 \longrightarrow \infty$. This means m_3 escapes from the system.

Recalling the general equation of motion

$$m_k \ddot{\underline{r}}_k = \sum_{1 \leq j < k \leq n} m_{jk} \frac{f(r_{jk})}{r_{jk}} (\underline{r}_j - \underline{r}_k) , \qquad (1)$$

when $f(r_{jk})$ is a real, analytic function there exists a unique set of $r_k(t)$s, containing the origin, which satisfies (1). For the case where $f(r) = \frac{1}{r^2}$, either all the $r_k(t)$s may be continued analytically as $t \longrightarrow \infty$, or there is some time $t = t_1$ at which at least one r_k ceases to be analytic. Painlevé has shown that the solution of the n - body problem permits analytic continuation until such time $t = t_1$ for which $r_k(t) \longrightarrow 0$ as $t \longrightarrow t_1$, and that this condition is both necessary and sufficient.

Returning to the work of Chazy for the case where $t \longrightarrow \infty$, is it possible to find estimates for the growth in r(t) with time? Consider

$$U = \sum \frac{m_{jk}}{r_{jk}} ,$$

$$-\dot{U} = \sum \frac{m_{jk} \dot{r}_{jk}}{r_{jk}^2} , \text{ but } \frac{1}{r_{jk}} \leq \frac{1}{r} \leq CU, \text{ so}$$

$$\left|\dot{U}\right| \leq CU^2 \sum m_{jk} \dot{r}_{jk} . \qquad (2)$$

As before, we square both sides of (2), recalling the energy equations, to get

$$\left|\dot{U}\right|^2 \leq CU^4 T, \qquad (3)$$

where $T = U + E > 0$.

If $E < 0$, from $\frac{1}{r} \leqslant CU$ and $T = U + E$, $r \leqslant B < \infty$. (Note that in all these lectures no effort has been made to distinguish between the various constants, i.e., A, B, C, C_1, etc., since they only depend on the masses). Thus we have

$$\frac{A}{r} \geqslant U \geqslant |E| \text{ , or } r \leqslant \frac{A}{|E|} \Longrightarrow r \sim t^0. \tag{4}$$

If $E = 0$, $T = U$ and (3) becomes

$$|\dot{U}|^2 \leqslant CU^5, \text{ or } \frac{|\dot{U}|}{U^{5/2}} \leqslant B. \tag{5}$$

Let $U = \frac{1}{\xi}$, differentiate, and substitute in (5) to get $\xi^{1/2} |\xi| \leqslant B$, $\xi^{3/2} \leqslant Ct$, but $r \leqslant C \xi$ from $\frac{A}{r} \geqslant U \quad \frac{1}{\xi}$, so

$$r \sim Ct^{2/3} \tag{6}$$

If $E > 0$, (3) becomes

$$|\dot{U}| \leqslant CU^4 (U + E). \tag{7}$$

Again let $U = \frac{1}{\xi}$, so (7) can be written in the form

$$\left| \frac{\dot{\xi}^2}{\xi^2} \right| \leqslant \frac{C}{\xi^4} \left(\frac{1}{\xi} + E \right) , \text{ or}$$

$$B \left| \dot{\xi} \right| \leqslant \sqrt{\frac{\xi}{1 + E\xi}} ,$$

$$\left| \dot{\xi} \right| \leqslant C .$$

Thus $\xi \leq Ct$ and $r \leq Ct$, so

$$r \sim t. \tag{8}$$

Since we have only worked with Lagrange - Jacobi equation to date, let us see what additional information can be obtained from these differential equations for the n - body case, where $E > 0$ and $\mathbb{1} \neq 0$. Making use of lemma 2 in the last lecture,

$$I \geq At^2 \Longrightarrow R \geq At, \text{ so } J \geq At.$$

But $\frac{1}{UJ} \geq \delta > 0$ if $\mathbb{1} \neq 0$, so that

$$\frac{1}{U} \geq \delta J \geq Ct \Longrightarrow U \leq \frac{C}{t}.$$

From $\ddot{\underline{r}}_k + \sum \frac{m_j}{r_{jk}^3} (\underline{r}_j - \underline{r}_k)$,

$$\left| \ddot{\underline{r}}_k \right| \leq \sum \frac{m_j}{r_{jk}^3} r_{jk} \leq \sum \frac{m_j}{r_{jk}^2} \leq AU^2 \leq \frac{B}{t^2}. \tag{9}$$

Summarizing Chazy's results for n = 3 and ruling out the case of triple collision, i.e., $\underline{h} \neq 0$, we see:

$E > 0$	all $r_{ij} \sim t$	$\mathbb{1} \neq 0$	hyperbolic case
	two r_{ij}s, say r_{12} and $r_{23} \sim t$, $r_{31} \sim t^{2/3}$	$\mathbb{1} = 0$	hyperbolic - parabolic case
	two r_{ij}s, say r_{12} and $r_{23} \longrightarrow \infty$, r_{12}, $r_{23} \sim t$, $r_{31} \leq B$ (bdd.)	$\mathbb{1} = 0$	hyperbolic - elliptic case

$E = 0$	same as above all $r_{ij} \sim t^{2/3}$	hyperbolic - elliptic, or parabolic case.
$E < 0$	two particles collide, the third $\longrightarrow \infty$	hyperbolic - elliptic, or parabolic - elliptic case.

Some r_{jk} is neither bounded nor unbounded.

Chazy was able to demonstrate orbits of every type above save for the very last, but the latest Russian literature indicates this too has now been demonstrated.

Let us now return to the Sundman problem, namely the three body problem where $h \neq 0$, and show that if $r \longrightarrow 0$ as $t \longrightarrow t_1$, this corresponds to a two particle collision with the third particle moving to a definite position with a definite velocity.

If $r(t) \longrightarrow 0$ as $t \longrightarrow t_1$, $U \longrightarrow \infty$ so $\ddot{I} \longrightarrow \infty$. But this tells us that the curve I vs. t must be concave upward in the neighborhood of $t = t_1$, so $I \longrightarrow L$, where $0 \leqslant L \leqslant \infty$. We have already ruled out the case where $I = 0$, so we have $0 < L \leqslant \infty$. Let us now show the 2 particle collision.

If $r \longrightarrow 0$, a pair of particles collide. Then one of the distances r_{ij} becomes a minimum and remains so. To prove this, assume r_{12} and r_{23} are alternately the minimum. Then there exists a sequence of times $\{t_n\}$, $t_n \longrightarrow t_1$ where $r_{12}(t_n) = r_{23}(t_n) \longrightarrow 0$. But $r_{31} \leqslant r_{12} + r_{23}$, so $r_{31} \longrightarrow 0$ which implies all three $r_{ij} \longrightarrow 0$ along this sequence of t_n's. But note that

$$I = \frac{1}{2M} \sum m_{jk} r_{jk}^2 \text{ , so}$$

$I(t_n) \longrightarrow 0$ along this sequence.

We have already proved I has a limit, so if it approaches 0 along a particular sequence of times, it will approach 0 no matter how you approach t_1. But this is a triple collision, contrary to $\underline{h} \neq 0$. Thus only one r_{ij} eventually becomes and stays the minimum.

Now to show that the third particle moves to a definite position with a definite velocity, return to the inequality $0 < L \leqslant \infty$, and let us show we can rule out $L = \infty$ when $\underline{h} \neq 0$. If $r_{12} \longrightarrow 0$, $r_{23} - r_{31} \longrightarrow 0$. (10)

$$I = \frac{1}{2M}\left\{ m_{12} r_{12}^2 + m_{23} r_{23}^2 + m_{31} r_{31}^2 \right\} \tag{11}$$

We know $m_{12} r_{12}^2 \longrightarrow 0$, so let us assume $L = \infty$, or $I \longrightarrow \infty$ as $t \longrightarrow t_1$. This implies

$$m_{23} r_{23}^2 + m_{31} r_{31}^2 \longrightarrow \infty .$$

Let $m_{23} > m_{31}$, so we may write

$$m_{23}(r_{23}^2 + r_{31}^2) \longrightarrow \infty \text{ or}$$

$$r_{23}^2 + r_{31}^2 \longrightarrow \infty .$$

From (10), if $r_{23} \longrightarrow \infty$, $r_{31} \longrightarrow \infty$ since

$r_{23} = r_{31} + 2\epsilon$, where $\epsilon \longrightarrow 0$ as $t \longrightarrow t_1$, and by squaring,

simplifying, and neglecting vanishing terms

$$r_{31} + \epsilon \longrightarrow 0 \Longrightarrow r_{31} \longrightarrow \infty \tag{12}$$

This implies $r_{23} \longrightarrow \infty$.

But $\ddot{\underline{r}}_3 = \frac{m_1}{r_{13}^3}(\underline{r}_1 - \underline{r}_3) + \frac{m_2}{r_{23}^3}(\underline{r}_2 - \underline{r}_3)$, so

$$\left|\ddot{\underline{r}}_3\right| \leq \frac{m_1}{r_{13}^2} + \frac{m_2}{r_{23}^2} \quad . \tag{13}$$

Using (12) in (13)

$$\left|\ddot{\underline{r}}_3\right| \longrightarrow 0 \text{ as } t \longrightarrow t_1 \text{ so}$$

$$\dot{\underline{r}}_3 \text{ has a limit as } t \longrightarrow t_1. \tag{14}$$

Recalling that if a function has a bounded derivative as $t \longrightarrow t_0$, the function itself has a limit, (14) implies

$$r_3 \longrightarrow \text{limit as } t \longrightarrow t_1 \text{ or}$$

$$\lim_{t \to t_1} r_3(t) = 1. \tag{15}$$

However, $m_1\underline{r}_1 + m_2\underline{r}_2 + m_3\underline{r}_3 = 0$, amd

$$m_1\underline{r}_1 + m_2\underline{r}_2 \longrightarrow L. \tag{16}$$

Also, $r_{12} \longrightarrow 0$, $\underline{r}_1 - \underline{r}_2 \longrightarrow 0$ (17)

Multiplying (17) by m_2 and adding to (16)

$$(m_1 + m_2)\, r_1 \longrightarrow L \Longrightarrow r_1,\ r_2 \text{ and } r_3 \text{ have limits.}$$

But from (11), since $I \longrightarrow L$ we have a contradiction to the assumption $L = \infty$.

$$\therefore\ 0 < L < \infty.$$

Summary of the theorems involved in the Sundman problem, where $n = 3$, $h \neq 0$, and $r(t) \longrightarrow 0$ as $t \longrightarrow t_1 < \infty$.

Theorem 1. $\lim_{t \to \infty} I(t) = L, \quad 0 < L < \infty$.

Theorem 2. Two particles collide and the third particle goes to a definite position with a finite velocity.

Theorem 3. If v_1 and v_2 are the velocities of the colliding particles relative to the origin, then

$$\left.\begin{aligned} \lim_{t \to t_1} r(t) v_1^2 &= 2m_2^2/\mu \\ \lim_{t \to t_1} r(t) v_2^2 &= 2m_1^2/\mu \end{aligned}\right\} \quad \mu = m_1 + m_2$$

Theorem 4. The integral

$$\int^{t} U(\tau)\, d\tau \text{, converges. (cf. } h = 0 \text{ in 2 body case).}$$

Theorem 5. If $\underline{V}$ is the velocity of m_1 relative to m_2, then at collision

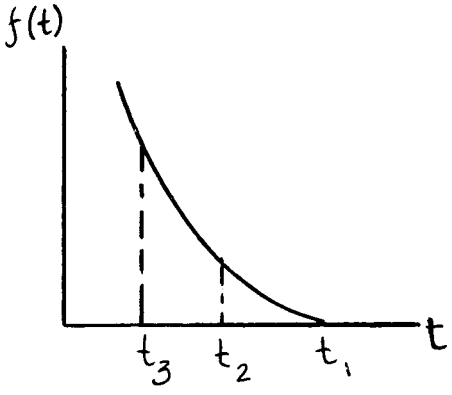
$$\lim_{t \to t_1} rV^2 = 2\mu . \qquad \mu = m_1 + m_2.$$

Theorem 6. $\lim_{t \to t_1} r(r^2)'' = 2\mu$

Theorem 7. $\lim_{t \to t_1} \dfrac{\underline{r}_2 - \underline{r}_1}{r_{12}}$ exists.

That is, the particles collide at a definite angle.

Proof that if $f(t) > 0$, $f(t_1) = 0$ and $f''(t) > 0$, then $f'(t) < 0$ in the neighborhood of $t = t_1$.



Let us expand $f(t)$ about $t = t_2$ in a Taylor series up to second order terms.

$$f(t_2) = f(t_3) + (t_2 - t_3)f'(t_3) + \frac{(t_2 - t_3)^2}{L^2} f''(\xi). \qquad (1)$$

$$t_2 \leq \xi \leq t_3$$

As $t_2 \longrightarrow t_1$, (1) becomes $0 \geqslant f(t_3) + (t_1 - t_3)f'(t_3)$, (2)

where we have neglected the second order term, which by hypothesis is positive. From (2)

$(t_1 - t_3)f'(t_3) < 0$. But $t_1 - t_3 > 0$, so $f'(t_3) < 0$.

Returning to Sundman's problem, consider the velocity of collision. From Theorem 7.2 in the 1962 lecture notes by Pollard, we have

$$rv_1^2 \longrightarrow \frac{2m_2^2}{m_1 + m_2},$$

$$rv_2^2 \longrightarrow \frac{2m_1^2}{m_1 + m_2} \text{ as } t \longrightarrow t_1.$$

Proof: Assume $r_{12} = r$, and recall that

$$U = \frac{m_{12}}{r_{12}} + \frac{m_{23}}{r_{23}} + \frac{m_{31}}{r_{31}} . \tag{3}$$

Multiply (3) by r and let $t \longrightarrow t_1$, to obtain

$$\lim_{t \to t_1} rU = m_{12}. \tag{4}$$

From (4)

$$r(t - E) \longrightarrow m_{12}.$$

But E is constant, $r \longrightarrow 0$, so we can write

$$rT \longrightarrow m_{12}, \text{ or } r(m_1v_1^2 + m_2v_2^2 + m_3v_3^2) \longrightarrow 2m_{12} \tag{5}$$

But $v_3 \longrightarrow \mathfrak{L} \Longrightarrow rv_3 \longrightarrow 0$, so (5) becomes

$$r(m_1v_1^2) + r(m_2v_2^2) \longrightarrow 2m_{12}. \tag{6}$$

Since the center of mass is such that

$$m_1v_1 + m_2v_2 + m_3v_3 = 0, \tag{7}$$

let us multiply (7) by $\sqrt{r}$, recall that $v_3 \longrightarrow 0$, $\sqrt{r} \longrightarrow 0$, and combine (6) and (7) to get

$$\begin{aligned} rv_1^2 &\longrightarrow \frac{2m_2^2}{m_1 + m_2} , \\ rv_2^2 &\longrightarrow \frac{2m_1^2}{m_1 + m_2} . \end{aligned} \tag{8}$$

Let us now find the rate at which the collision takes place.

Even though $r \longrightarrow 0$, the $\lim_{t \to t_1} \int^{t_1} \frac{d\tau}{r(\tau)}$ is bounded.

Since $\frac{1}{r} \sim U = \ddot{I} - 2E$, $\ddot{I} \longrightarrow \infty$ as $t \longrightarrow t_1$, so at some point $\ddot{I}$ becomes positive. Thus, if we show the existence of a limit for $\int^{t_1} \ddot{I}\, d\tau$, we have a limit for $\int^{t_1} \frac{d\tau}{r(\tau)}$.

Now,
$$I = m_1 r_1^2 + m_2 r_2^2 + m_3 r_3^2 ,$$
$$\dot{I} = m_1(\underline{r}_1 \cdot \underline{v}_1) + m_2(\underline{r}_2 \cdot \underline{v}_2) + m_3(\underline{r}_3 \cdot \underline{v}_3) .$$

Since r_3 and v_3 have limits,

$m_1 v_1 + m_2 v_2 \longrightarrow 0$ implies

$$m_2(\underline{r}_2 - \underline{r}_1) \cdot v_2 \leq m_2 r v_2 \equiv m_2 \sqrt{r}\sqrt{r}\, v_2 .$$

But $\sqrt{r}\, v_2$ has a limit from (8), $\sqrt{r}$ is likewise bounded. Let $V = \underline{v}_2 - \underline{v}_1$, and combining results, we see that

$$r v^2 \longrightarrow 2(m_1 + m_2)$$

Proof that $r(r^2)^{\cdot\cdot} \longrightarrow 2(m_1 + m_2)$, where r^2 is the square of the vector, and $\underline{r} = \underline{r}_2 - \underline{r}_1$.

$$\frac{d^2}{dt^2} r(r^2) = 2r\left\{ \underline{r} \cdot \ddot{\underline{r}} + \dot{\underline{r}}^2 \right\} \qquad (9)$$

$$\ddot{\underline{r}}_1 = \frac{m_2}{r_{12}^3} (\underline{r}_2 - \underline{r}_1) + \text{bdd. fcn.} \tag{10}$$

$$\ddot{\underline{r}}_2 = \frac{m_1}{r_{12}^3} (\underline{r}_1 - \underline{r}_2) + \text{bdd. fcn.} \tag{11}$$

Subtract (10) from (11).

$$\ddot{\underline{r}} = - \frac{m_1 + m_2}{r^3} \underline{r} + \text{bdd. fcn.}$$

So, $\underline{r} \cdot \ddot{\underline{r}} = - \frac{m_1 + m_2}{r} + \text{bdd. fcn.},$

$$2r(\underline{r} \cdot \ddot{\underline{r}}) = -2(m_1 + m_2) + \text{bdd. fcn.} \tag{12}$$

From (9) and (12)

$$2r\left\{\underline{r} \cdot \ddot{\underline{r}} + \dot{\underline{r}}^2\right\} = - 2(m_1 + m_2) + \text{bdd. fcn.} + 4(m_1 + m_2) + \text{vanishing term.}$$

$$\therefore \quad r(r^2)^{\cdot\cdot} \longrightarrow 2(m_1 + m_2) \tag{13}$$

Now use this information to find the rate at which $r \longrightarrow 0$ as $t \longrightarrow t_1$. In the two-body problem $r \sim (t_1 - t)^{2/3}$. Can the same be said of the three-body problem?

Differentiating $r(r^2)^{\cdot\cdot}$ as a scalar function, we get

$$2r(r\ddot{r} + \dot{r}^2) \ .$$

Define $F = \frac{r^2 \dot{r}^2}{r}$, where both numerator and denominator $\longrightarrow 0$ as $t \longrightarrow t_1$. Applying L'Hôpital's rule,

$$\lim_{t\to t_1} F = \lim_{t\to t_1} \frac{r^2(2\dot{r}\ddot{r}) + \dot{r}^2(2r\dot{r})}{\dot{r}} = \lim_{t\to t_1} 2r(\dot{r}^2 + r\ddot{r})$$

$$\therefore \lim_{t\to t_1} F = 2(m_1 + m_2).$$

Thus $r\dot{r}^2 \longrightarrow 2(m_1 + m_2)$, or

$$\sqrt{r}\,\dot{r} \longrightarrow \pm\sqrt{2(m_1 + m_2)} \,. \tag{14}$$

But $r\to 0$, so (14) becomes

$$\sqrt{r}\,\dot{r} \longrightarrow -\sqrt{2(m_1 + m_2)}\,, \text{ and}$$

$$\int_r^0 \sqrt{r}\,dr \sim -\sqrt{2(m_1 + m_2)}\ (t_1 - t). \tag{15}$$

Integrating the left side of (15)

$$\frac{2}{3} r^{3/2} \sim \sqrt{2(m_1 + m_2)}\ \ (t_1 - t) \Longrightarrow r \sim (t_1 - t)^{2/3} \tag{16}$$

OUTLINE OF A THEORY OF NON-PERIODIC MOTIONS IN THE NEIGHBORHOOD OF THE LONG-PERIOD LIBRATIONS ABOUT THE EQUILATERAL POINTS OF THE RESTRICTED PROBLEM OF THREE BODIES

by

Professor E. Rabe
University of Cincinnati Observatory
Cincinnati, Ohio

I. Introduction.

Subsequent to the recent determination of selected periodic libration orbits of the Trojan type (Rabe, 1961, 1962) and of the corresponding orbits in the restricted earth-moon problem (Rabe and Schanzle, 1962), additional numerical work has been devoted to the study of non-periodic trajectories, i.e., orbits deviating from a certain periodic solution by given initial quantities. Such orbit computations have been limited to the Trojan problem because of the greater simplicity of these motions which are based on the relatively small mass ratio Jupiter/sun. Also the work has been limited to motions in the plane of the periodic orbits, and consequently in the plane of Jupiter's orbit. A number of selected non-periodic orbits of the Trojan type were thus obtained in cooperation with J. Schubart, during the summer of 1962, on the SIEMENS-2002 electronic computer at the Astronomisches Rechen-Institut in Heidelberg. An even more extensive and systematic survey of the possible forms of motion was undertaken by A. Schanzle (Dissertation in preparation) on the I.B.M.-1620 of the University of Cincinnati, for a study of the stability characteristics.

From the rather numerous trajectories, some of which extend over time intervals of many hundred years, the following principal features emerged rather clearly. For very small initial deviations from a periodic solution with the same value of the Jacobi constant C as the non-periodic orbit, the latter oscillates in a vine-like fashion about the periodic reference orbit, with a principal short period of the order of Jupiter's period of orbital revolution (as compared to the roughly 13 times longer period of libration). For increasingly larger initial displacements, however,

the non-periodic orbit detaches itself from the periodic one with the same C-value, first in the regions of the two turning points, describing complete short-period loops outside of the periodic orbit and increasing in this manner the effective libration amplitude of the non-periodic trajectory. The over-all librational behavior is maintained, but with an increased amplitude depending on the initial deviation from the periodic orbit. No indication of real instability is observed, even when the dimensions of the non-periodic "libration" are many times those of the related periodic orbit. Clearly, however, the concept of "ordinary stability" does not cover such phenomena, but a special concept of "librational stability" may be appropriate. It may be characterized and defined by one feature which seems to apply to all the observed non-periodic librational motions: The appearance of ordinary stability (vine-like oscillations about a reference orbit, without any closed loops outside of it) can be restored by referring the non-periodic trajectory to some other periodic libration orbit of larger amplitude and with a correspondingly larger Jacobi constant C_0, instead of to the one with the value C of the non-periodic orbit.

The observed features are those in the rotating coordinate system of the restricted problem. If the related heliocentric osculating orbital elements are computed at various points of the librational motion, then it is found that during the whole libration the semi-major axis a of the non-periodic Trojan follows rather closely the long-period fluctuation of the periodic Trojan in that libration orbit which approximates best the amplitude of the non-periodic libration. The eccentricity e, however,

fluctuates very little and is roughly proportional to the amplitude of the principal short-period oscillation in the rotating frame of reference. In this connection it should be noted that the eccentricity of the periodic Trojan is always very small, or, more precisely, of the order of the mass ratio μ of the two finite masses. The observed behavior of the elements a and e can be understood on the basis of the Tisserand criterion

$$\frac{1}{a} + 2\left[a(1 - e^2)\right]^{\frac{1}{2}} \approx \frac{C}{1+\mu} , \qquad (1)$$

as an approximate equivalent of the Jacobi integral. With Jupiter's solar distance a' = 1 as the unit of length, Eq. (1) shows that near a = 1 the required constancy of the left-hand side is indeed compatible with rather substantial variation of a, in combination with much smaller variations of e.

Considering the stability suggested by the numerical results as described above, it appears desirable to attempt an analytical representation of these non-periodic motions in the form of oscillation terms of various periods superimposed on a periodic libration as reference or intermediate orbit. In the Trojan problem as well as in the earth-moon case, such reference orbits are available in the form of their Fourier series representations, and for the Trojans they can be interpolated between the directly computed periodic orbits (Rabe, 1962) to find any desired periodic solution within the amplitude range of the real Trojan planets.

In the theory thus proposed, the basic periodic

orbit will evidently play the same role as Hill's variation orbit in the lunar theory. In further analogy, the non-periodic Trojan's oscillation about the reference orbit is approximately proportional to the eccentricity in its heliocentric orbit, just as the moon's deviation from the lunar variation orbit is related to its orbital eccentricity. Here, however, the analogy ends, because the intermediate Trojan orbit differs essentially, geometrically as well as analytically, from Hill's variation orbit. The sharp curvature and small velocity near the two turning points of the libration orbit in particular are quite unique features of the problem at hand. This phenomenon impairs also the usefulness of the well known second order differential equation of Hill for the normal displacement from the reference orbit, and for the discussion of the stability of the latter. It had been found that the periodic function f(u) in Hill's equation (Rabe, 1961) converges rather poorly in the case of the periodic Trojan orbits, and has sharp and deep dips at the two turning points. This behavior can also be understood on the basis of the computed non-periodic orbits, as well as in the light of the analytical results to be presented here. To the first order of approximation, the principal short-period oscillation superimposed on the long-period libration orbit has the shape of an ellipse with a 2/1 ratio of its principal axes, and the large axis tends to keep itself aligned with the tangent to Jupiter's orbit, but not with the tangent to the libration orbit in the rotating frame. In other words, the short-period oscillations do not follow the curvature of the libration orbit and consequently cannot be represented by just a few terms from a solution of Hill's equation.

II. The Differential Equations for the Displacements from the Periodic Libration Orbit.

If the origin of the rotating rectangular coordinate system (x,y) is identified with the center of mass of the two finite masses, both of which are permanently at rest on the x-axis, then the periodic librations are known functions

$$x = x_0 + \sum_{j=1}^{\infty} x_{c,j} \cos(j\sigma) + \sum_{j=1}^{\infty} x_{s,j} \sin(j\sigma)$$
$$y = y_0 + \sum_{j=1}^{\infty} y_{c,j} \cos(j\sigma) + \sum_{j=1}^{\infty} y_{s,j} \sin(j\sigma), \qquad (2)$$

with given numerical coefficients x_0, y_0, $x_{c,j}$, $x_{s,j}$, $y_{c,j}$, $y_{s,j}$, and with

$$\sigma = \frac{2\pi}{T}(t - t_0) = n(t - t_0). \qquad (3)$$

In Eq. (3), T is the period of the libration, and t_0 denotes the moment when the periodic Trojan planet intersects the straight line connecting the principal mass (sun or earth, respectively) with the triangular point L_5, at the outside passage of L_5. The quantity n, also defined by Eq. (3), may be called the frequency or mean motion of the libration. In Eqs. (2), only the coefficient x_0 differs from the corresponding p_0 in the earlier (p,q)-system (Rabe, 1961), in consequence of the different origin of the (x,y)-system. The convergence of the series represented by Eqs. (2) was found to be very satisfactory for a wide range of libration amplitudes, in the sun-Jupiter as well as in the earth-moon case. The theory of non-periodic motions to be outlined here applies to both cases,

but for the sake of a convenient terminology everything will be phrased in terms of the Trojan problem, for which it is also intended to use the results first of all.

The periodic solution (2) satisfies the differential equations

$$\begin{aligned} \ddot{x} - 2N\dot{y} &= \Omega x \\ \ddot{y} + 2N\dot{x} &= \Omega y \ , \end{aligned} \tag{4}$$

where Ωx and Ωy denote the partial derivatives with respect to x and y of the function

$$\Omega = \frac{1}{\Delta_1} + \frac{\mu}{\Delta_2} + \frac{1}{2}\left(\Delta_1^2 + \mu\Delta_2^2\right) . \tag{5}$$

Eqs. (4) are based on the adoption of the constant distance sun-Jupiter as the unit of length, of the mass of the sun as the unit of mass, and of a unit of time which reduces the gravitational constant to unity. Therefore, Jupiter's angular orbital velocity N is related to its mass μ by

$$N^2 = 1 + \mu \ . \tag{6}$$

In Eqs. (5) for Ω, the quantities Δ_1 and Δ_2 denote the periodic Trojan's distance from the sun and from Jupiter, respectively, and are given by

$$\begin{aligned} \Delta_1^2 &= \left(x - \frac{\mu}{1+\mu}\right)^2 + y^2 \\ \Delta_2^2 &= \left(x + \frac{1}{1+\mu}\right)^2 + y^2 \ . \end{aligned} \tag{7}$$

Eqs. (7) indicate that the sun is permanently located at $\left(+\frac{\mu}{1+\mu}\right)$, 0 , and Jupiter at $\left(-\frac{1}{1+\mu}\right)$, 0 .

The differential equations (4) have to be satisfied not only by the particular, periodic solutions (2), but also by any

$$\bar{x} = x + u \qquad \bar{y} = y + v \tag{8}$$

representing the motion of a non-periodic Trojan oscillating, by increments (u,v), about the (x,y) of the periodic or reference solution (2). It is easily seen that the replacement of x,y by x+u, y+v in Eqs. (4), and the subsequent subtraction of the original Eqs. (4), produces the new differential equations for u,v in the form

$$\begin{aligned} \ddot{u} - 2N\dot{v} &= \Omega_{xx}u + \Omega_{xy}v + \tfrac{1}{2}\Omega_{xxx}u^2 + \tfrac{1}{2}\Omega_{xyy}v^2 + \Omega_{xxy}uv + \dots \\ \ddot{v} + 2N\dot{u} &= \Omega_{xy}u + \Omega_{yy}v + \tfrac{1}{2}\Omega_{xxy}u^2 + \tfrac{1}{2}\Omega_{yyy}v^2 + \Omega_{xyy}uv + \dots , \end{aligned} \tag{9}$$

where the $\Omega_x(x+u, y+v)$ and $\Omega_y(x+u, y+v)$ originally involved on the right-hand sides have been expanded as Taylor series in powers of u,v. The Ω_{xx}, Ω_{xy}, etc., denote the second and higher order partials of Ω with respect to x and y, as functions of x and y alone (with $u = v = 0$), and thus as periodic functions of σ .

Two different possibilities exist for the determination of the Ω_{xx}, Ω_{xy}, etc., on the basis of the corresponding periodic solution (2). First, differentiation of Eqs. (4) with respect to the time t produces

$$\ddot{x} - 2N\dot{y} = \Omega_{xx}\dot{x} + \Omega_{xy}\dot{y}$$
$$\ddot{y} + 2N\dot{x} = \Omega_{xy}\dot{x} + \Omega_{yy}\dot{y} \ , \qquad (10)$$

two relations involving the three second order partials of Ω together with the known periodic functions $\dot{x}$, $\dot{y}$, $\ddot{x}$, $\ddot{y}$, $\dddot{x}$, $\dddot{y}$. The third relation still needed for the determination of Ω_{xx}, Ω_{yy}, Ω_{xy} is obtained from

$$\Omega_{xx} + \Omega_{yy} + \Omega_{zz} = 2N^2 \ , \qquad (11)$$

taking advantage of the fact that for motions limited to the (x,y)-plane, with $z \equiv 0$, one has

$$\Omega_{zz} = -\frac{1}{\Delta_1^3} - \frac{\mu}{\Delta_2^3} = -N^2 + \frac{1}{y}(2N\dot{x} + \ddot{y}) \ , \qquad (12)$$

so that

$$\Omega_{xx} + \Omega_{yy} = 3N^2 - \frac{1}{y}(2N\dot{x} + \ddot{y}) \ . \qquad (13)$$

Now Eqs. (10) and (13) may be solved for Ω_{xx}, Ω_{yy}, Ω_{xy}, and the four third order partials of Ω may then be found from relations obtained by differentiating Eqs. (10) and (13) with respect to t, and Eq. (13) also with respect to y. The continuation of these differentiation procedures evidently produces the necessary number of relations for the determination of the partials of Ω of any required order as periodic functions of σ. The solution of these systems of equations will involve divisions by various powers of y and of $V^2 = \dot{x}^2 + \dot{y}^2$, but for

the libration orbits considered here, these two quantities never vanish.

The second possibility for the determination of the functions Ω_{xx}, etc., is based on the general expressions for these partials in terms of x and y. These are easily derived from the corresponding differentiations of Eq. (5) for Ω, considering also the Eqs. (7) for Δ_1 and Δ_2. One finds

$$\Omega_{xx} = \left(1 - \frac{1}{\Delta_1^3}\right) + \frac{3}{\Delta_1^5}\left(x - \frac{\mu}{1+\mu}\right)^2 + \mu\left[\left(1 - \frac{1}{\Delta_2^3}\right) + \frac{3}{\Delta_2^5}\left(x + \frac{1}{1+\mu}\right)^2\right]$$

$$\Omega_{yy} = \left(1 - \frac{1}{\Delta_1^3}\right) + \frac{3}{\Delta_1^5} y^2 + \mu\left[\left(1 - \frac{1}{\Delta_2^3}\right) + \frac{3}{\Delta_2^5} y^2\right] \qquad (14)$$

$$\Omega_{xy} = \frac{3}{\Delta_1^5}\left(x - \frac{\mu}{1+\mu}\right) y + \mu \frac{3}{\Delta_2^5}\left(x + \frac{1}{1+\mu}\right) y \quad ,$$

$$\Omega_{xxx} = \frac{9}{\Delta_1^5}\left(x - \frac{\mu}{1+\mu}\right) - \frac{15}{\Delta_1^7}\left(x - \frac{\mu}{1+\mu}\right)^3 + \mu\left[\frac{9}{\Delta_2^5}\left(x + \frac{1}{1+\mu}\right) - \frac{15}{\Delta_2^7}\left(x + \frac{1}{1+\mu}\right)^3\right]$$

$$\Omega_{yyy} = \frac{9}{\Delta_1^5} y - \frac{15}{\Delta_1^7} y^3 + \mu\left[\frac{9}{\Delta_2^5} y - \frac{15}{\Delta_2^7} y^3\right]$$

$$\Omega_{xxy} = \frac{3}{\Delta_1^5} y - \frac{15}{\Delta_1^7}\left(x - \frac{\mu}{1+\mu}\right)^2 y + \mu\left[\frac{3}{\Delta_2^5} y - \frac{15}{\Delta_2^7}\left(x + \frac{1}{1+\mu}\right)^2 y\right] \qquad (15)$$

$$\Omega_{xyy} = \frac{3}{\Delta_1^5}\left(x - \frac{\mu}{1+\mu}\right) - \frac{15}{\Delta_1^7}\left(x - \frac{\mu}{1+\mu}\right) y^2 + \mu\left[\frac{3}{\Delta_2^5}\left(x + \frac{1}{1+\mu}\right) - \frac{15}{\Delta_2^7}\left(x + \frac{1}{1+\mu}\right) y^2\right],$$

etc. Evidently these expressions offer no advantage over the Eqs. (10) and (13) and those following from the differentiation of (10) and (13) if a direct substitution of the periodic functions x,y is contemplated. However, the Eqs. (14), (15), etc., are very convenient for the determination of all these partials of Ω by harmonic analysis, which is facilitated by the availability of the special values of x and y for all the periodic orbits obtained in the Trojan case as well as in the earth-moon system. For interpolated periodic Trojan orbits, the required special values of x and y, for equidistant values of σ, may be obtained either by interpolation between the

special values of the actually computed orbits, or by interpolation of the coefficients x_0, y_0, $x_{c,1}$, $x_{s,1}$, etc., and subsequent computation of the needed values of x and y from Eqs. (2). Since the Ω_{xx}, Ω_{yy}, etc., are the coefficients of the displacements u,v and of their various powers in the differential equations (9), a reduced accuracy will be permitted, depending on the amplitude of the oscillations represented by u and v.

Once the Fourier expansions of the partials of Ω have been obtained by either of the two methods, we may assume to possess all of them in the form now given for Ω_{xx}, Ω_{yy}, and Ω_{xy}:

$$\Omega_{xx} = A_0 + 2 \sum_{r=1}^{\infty} \left[A_{c,r} \cos (r\sigma) + A_{s,r} \sin (r\sigma) \right]$$

$$\Omega_{yy} = B_0 + 2 \sum_{r=1}^{\infty} \left[B_{c,r} \cos (r\sigma) + B_{s,r} \sin (r\sigma) \right] \qquad (16)$$

$$\Omega_{xy} = C_0 + 2 \sum_{r=1}^{\infty} \left[C_{c,r} \cos (r\sigma) + C_{s,r} \sin (r\sigma) \right] .$$

As far as the size of the various coefficients is concerned, it is clear that for <u>small</u> librations the A_0, B_0, C_0 should not differ very much from the values of Ω_{xx}, Ω_{yy}, Ω_{xy} at the libration point L_5, which are well known, and also easily obtained from Eqs. (14) as follows:

$$(\Omega_{xx})_0 = + \tfrac{3}{4} (1 + \mu), \quad (\Omega_{yy})_0 = + \tfrac{9}{4} (1 + \mu),$$

$$(\Omega_{xy})_0 = - \tfrac{3}{4}\sqrt{3} (1 - \mu). \qquad (17)$$

The corresponding values of the third order partials, at

L_5, are similarly obtained from Eqs. (15):

$$(\Omega_{xxx})_0 = -\frac{21}{8}(1-\mu)\,, \qquad (\Omega_{yyy})_0 = -\frac{9}{8}\sqrt{3}\,(1+\mu)$$
$$(\Omega_{xxy})_0 = -\frac{3}{8}\sqrt{3}\,(1+\mu)\,, \qquad (\Omega_{xyy})_0 = +\frac{33}{8}(1-\mu)\,. \tag{18}$$

These values are listed here for subsequent reference. The principal periodic terms in Eqs. (16), with subscripts c,1 and s,1, should be of the order of the coefficients $x_{c,1}$, $x_{s,1}$, $y_{c,1}$, $y_{s,1}$ in Eqs. (2), because Eqs. (14) indicate that the differences

$$\Omega_{xx} - A_0, \quad \Omega_{yy} - B_0, \quad \Omega_{xy} - C_0$$

must be of the general order of the periodic parts of x and y, or of

$$x - x_0, \qquad y - y_0\,.$$

Moreover, the convergence of the series represented by Eqs. (16) may be expected, by the same way of reasoning, to be just as satisfactory as that of Eqs. (2) for the periodic solution itself.

It will be convenient to transform the series expansions of the Eqs. (16) into the exponential form. For this purpose, let

$$\begin{aligned}
&\alpha_0 = A_0, && \beta_0 = B_0, && \gamma_0 = C_0,\\
&\alpha_r = A_{c,r}+iA_{s,r}, && \beta_r = B_{c,r}+iB_{s,r}, && \gamma_r = C_{c,r}+iC_{s,r} \quad \text{for } r<0,\\
&\alpha_r = A_{c,r}-iA_{s,r}, && \beta_r = B_{c,r}-iB_{s,r}, && \gamma_r = C_{c,r}-iC_{s,r} \quad \text{for } r>0,\\
&i = \sqrt{-1}
\end{aligned} \tag{19}$$

,

so that

$$\Omega_{xx} = \sum_{-\infty}^{\infty} \alpha_r \exp(ir\sigma), \quad \Omega_{yy} = \sum_{-\infty}^{\infty} \beta_r \exp(ir\sigma),$$
$$\Omega_{xy} = \sum_{-\infty}^{\infty} \gamma_r \exp(ir\sigma). \tag{20}$$

III. The Principal Features of the Solution.

Considering at first only all those terms which are linear in u and v on the right-hand sides of the differential equations (9), with the coefficients now given by Eqs. (20), the solution may be assumed in the form

$$u = \sum_{-\infty}^{\infty} u_r \exp\left[i(r+c)\,\sigma\right], \qquad v = \sum_{-\infty}^{\infty} v_r \exp\left[i(r+c)\,\sigma\right]. \tag{21}$$

Just as in the case of Hill's equation, the unknown coefficients u_r, v_r and the stability exponent c have to be determined from the identities resulting from the substitution of Eqs. (21) into (9). Since each of the two Eqs. (9) furnishes one identity in the form of an infinite series for each value of the integer r, the following pair of equations of this kind has to be satisfied for each possible value of r in the left-hand terms:

$$\left[n^2(r+c)^2 + \alpha_0\right] u_r + \left[2Nn(r+c)i + \gamma_0\right] v_r$$
$$= -\sum_{s=1}^{\infty} \left[\alpha_s u_{r-s} + \alpha_{-s} u_{r+s} + \gamma_s v_{r-s} + \gamma_{-s} v_{r+s}\right]$$
$$-\left[2Nn(r+c)i - \gamma_0\right] u_r + \left[n^2(r+c)^2 + \beta_0\right] v_r$$
$$= -\sum_{s=1}^{\infty} \left[\gamma_s u_{r-s} + \gamma_{-s} u_{r+s} + \beta_s v_{r-s} + \beta_{-s} v_{r+s}\right]$$
$$r = 0, \pm 1, \pm 2, \ldots. \tag{22}$$

For small basic librations, i.e., for small coefficients $\alpha_1, \alpha_{-1}, \ldots, \gamma_{-1}$ in Eqs. (20), a rather good first approximation to the solution of the infinite system of equations (22) may be obtained by neglecting all α_r, β_r, γ_r except α_0, β_0, γ_0, and all u_r, v_r except u_0, v_0. The resulting two equations,

$$\begin{aligned}(n^2c^2 + \alpha_0)u_0 + (2Nnci + \gamma_0)v_0 &= 0\\ -(2Nnci - \gamma_0)u_0 + (n^2c^2 + \beta_0)v_0 &= 0,\end{aligned} \tag{23}$$

have a solution if c satisfies the condition

$$(n^2c^2 + \alpha_0)(n^2c^2 + \beta_0) - 4N^2n^2c^2 - \gamma_0^2 = 0 . \tag{24}$$

The Eqs. (23) and (24) become identical with the well known corresponding equations for the first approximation to the short-periodic solutions of infinitesimal dimensions about the libration point L_5 if, as justified for such small libration, one puts

$$\alpha_0 = (\Omega_{xx})_0 , \qquad \beta_0 = (\Omega_{yy})_0 , \qquad \gamma_0 = (\Omega_{xy})_0 , \tag{25}$$

with the $(\Omega_{xx})_0$ etc., as given in Eqs. (17). Then Eq. (24) is satisfied by a short-period frequency $\nu = nc$, approximated by

$$nc \approx 1 - \frac{23}{4}\mu . \tag{26}$$

Since $n = 2\pi/T$, the stability exponent c is given by

$$c \approx \frac{T}{2\pi} 1\left(-\frac{23}{4}\mu\right) , \tag{27}$$

and is of the order of 12.

With the proper result for c, either one of the two equations (23) may now be used to express u_0 in terms of v_0, or vice versa, and either u_0 or v_0 may be considered as arbitrary. Going back to trigonometric functions and real variables, the result of the first approximation can be written in the form

$$\begin{aligned} u &= 2u_{c,0} \cos(c\sigma) + 2u_{s,0} \sin(c\sigma) \\ v &= 2v_{c,0} \cos(c\sigma) + 2v_{s,0} \sin(c\sigma) , \end{aligned} \qquad (28)$$

where either $u_{c,0}$, $u_{s,0}$ or $v_{c,0}$, $v_{s,0}$ are arbitrary, and the remaining two coefficients depend on the two arbitrary ones and on c through relations equivalent to the complex equations (23) and (24). These relations between the coefficients of the solution (28) involve α_0, β_0, γ_0, which vary with the amplitude of the basic libration of long period, and which for large librations may differ substantially from the approximations (25). Consequently the elliptic approximation (28) to u,v is not simply the transposed short-period solution about L_5, but contains already the effect of the constant parts of Ω_{xx}, Ω_{yy}, and Ω_{xy}.

To improve the solution to include the effect of the principal periodic terms of the second order partials of Ω, the Eqs. (22) for $r = \pm 1$ have to be solved simultaneously with those for $r = 0$, considering the involvement of α_1, α_{-1}, ..., γ_{-1} on the right-hand sides. The resulting equations are

$$
\begin{aligned}
&\left[n^2(c-1)^2 + \alpha_0\right]u_{-1} + \left[2Nn(c-1)i + \gamma_0\right]v_{-1} + \alpha_{-1}u_0 + \gamma_{-1}v_0 = 0\\
&-\left[2Nn(c-1)i - \gamma_0\right]u_{-1} + \left[n^2(c-1)^2 + \beta_0\right]v_{-1} + \gamma_{-1}u_0 + \beta_{-1}v_0 = 0\\
&\alpha_1 u_{-1} + \gamma_1 v_{-1} + \left[n^2c^2 + \alpha_0\right]u_0 + \left[2Nnci + \gamma_0\right]v_0 + \alpha_{-1}u_1 + \gamma_{-1}v_1 = 0\\
&\gamma_1 u_{-1} + \beta_1 v_{-1} - \left[2Nnci - \gamma_0\right]u_0 + \left[n^2c^2 + \beta_0\right]v_0 + \gamma_{-1}u_1 + \beta_{-1}v_1 = 0\\
&\alpha_1 u_0 + \gamma_1 v_0 + \left[n^2(c+1)^2 + \alpha_0\right]u_1 + \left[2Nn(c+1)i + \gamma_0\right]v_1 = 0\\
&\gamma_1 u_0 + \beta_1 v_0 - \left[2Nn(c+1)i - \gamma_0\right]u_1 + \left[n^2(c+1)^2 + \beta_0\right]v_1 = 0 \,.
\end{aligned}
\tag{29}
$$

The improved value of c will result from the condition that the determinant of the coefficients of the six unknowns u_{-1}, v_{-1}, ..., v_1 has to vanish. Again, either u_0 or v_0 may be assumed as arbitrary, and the other five unknowns will be obtained from the remaining five independent equations of the system (29). On the basis of the experience with successive approximations to the corresponding solution of Hill's equation for the oscillations about the periodic Trojan orbits (Rabe, 1961, 1962), and of the numerical results for many non-periodic trajectories, it can be expected that even for relatively large librations the result from Eqs. (29) for c will not differ much from the first approximation, as obtained from Eq. (24), and that the coefficients u_{-1}, v_{-1}, u_1, v_1 will be found to be substantially smaller than u_0, v_0. The good convergence of the coefficients α_r, β_r, γ_r, as compared to the very poor convergence of the comparable θ_r when Hill's

equation is used (Rabe, 1961), should benefit not only the successive improvements of c, but also the successive determinations of the various u_r, v_r.

Up to this point, the second powers of u and v have been neglected in Eqs. (9). Let it be assumed that the solution considering the linear involvement of u and v, in the form of Eqs. (21), has been completed to the desired degree of accuracy, as far as the numerical size of the coefficients u_r, v_r is concerned. An addition to this solution to account for the actual presence of the second order terms in u and v in Eqs. (9) can obviously be obtained by substitution of the "linear solution" (21) into the previously neglected second order terms of Eqs. (9) and by the subsequent determination of a new solution satisfying these newly created terms on the right-hand sides (and, of course, the linear terms as well). If then the new solution is added to the previous one, the general solution of the Eqs. (9) will have been completed to the terms of the second order in u and v. After this, of course, a further "third order addition", considering the presence of cubic terms in u and v in the Eqs. (9), may be obtained in the same general manner, if necessary or desired. Attention has to be paid to the form of the exponents (arguments) emerging in each addition to the previously established solution, because the re-appearance of previously obtained exponents in such an addition may require a slight revision of the earlier solution for these terms.

The substitution of Eqs. (21) into the second order terms of Eqs. (9) will create exponents of the forms

$i(2c+r)\sigma$ and $ir\sigma$, including cases with $r = 0$. Accordingly, the additive solution should provide for all these exponents, none of which is of the form of those in the linear solution (21). The coefficients of the new terms will be determined by substitution into the complete Eqs. (9). The terms with exponents $i(2c+r)\sigma$ are of short period, those with $ir\sigma$, however, of long period, and constant terms appear when $r = 0$ in the latter group. The appearance of small constant terms in the right-hand sides of Eqs. (9) is easily absorbed by corresponding constant terms u_{00} and v_{00} in u and v, respectively. If the constants appearing in the two Eqs. (9) are k and λ, respectively, then u_{00} and v_{00} will be determined from

$$\begin{aligned} \alpha_0 u_{00} + \gamma_0 v_{00} &= -K \\ \gamma_0 u_{00} + \beta_0 v_{00} &= -\lambda \ . \end{aligned} \qquad (30)$$

If one proceeds to the consideration of terms involving u^3, u^2v, etc., in Eqs. (9), then it is seen that the substitution of the preceeding linear and second order solutions into these terms will partly produce new terms with exponents of the form $i(r+c)\sigma$ of the linear solution (21). This will require a slight revision of the earlier solution for these terms, including a corresponding refinement of the determination of c. No principal difficulty, however, appears to stand in the way of an extension of the total solution to any desired degree of precision. The purely periodic (or constant) nature of all the terms emerging in the process of the solution, combined with the very good convergence of the Fourier expansions involved, indicates true orbital stability, not just in the first order sense usually decided on the basis of the linear

terms alone, but considering the presence of the second and higher order terms as well. There is no indication of instability even for rather substantial amplitudes of u and v. These analytical results confirm the tentative conclusions from the numerous calculations of such non-periodic orbits.

IV. Some Properties of the Solution.

If V is the non-periodic Trojan's velocity in the rotating frame, and V_0 that of the periodic reference Trojan in the libration orbit, then one has at any time

$$V^2 - V_0^2 = 2(\dot{x}\dot{u} + \dot{y}\dot{v}) + \dot{u}^2 + \dot{v}^2 \ . \qquad (31)$$

Limiting x and y to their principal periodic terms with j = 1 in Eqs. (2), and u and v similarly to the approximation represented by Eqs. (28), the products $\dot{x}\dot{u}$ and $\dot{y}\dot{v}$ involved in Eq. (31) are found to consist of periodic terms only, with arguments $(c\pm1)\sigma$. The part $\dot{u}^2+\dot{v}^2$, however, contributes a constant part

$$\overline{V^2 - V_0^2} = 2c^2u^2(u_{c,0}^2 + u_{s,0}^2 + v_{c,0}^2 + v_{s,0}^2) \ . \qquad (32)$$

The difference $\Omega - \Omega_0$ between the functions Ω and Ω_0 associated with the non-periodic and the periodic Trojan, respectively, can be expanded in the form

$$\Omega - \Omega_0 = \Omega_x u + \Omega_y v + \tfrac{1}{2}\Omega_{xx}u^2 + \tfrac{1}{2}\Omega_{yy}v^2 + \Omega_{xy}uv + \ldots . \qquad (33)$$

No constant terms arise from the linear part of Eq. (33), but the second order part contributes such terms. From the approximations of Eqs. (17), for Ω_{xx} etc., and from Eqs. (28) for u and v, the constant part of $\Omega - \Omega_0$ is found to be of the form

$$\overline{\Omega - \Omega_0} = \frac{3}{4}(1+\mu)(u_{c,0}^2 + u_{s,0}^2) + \frac{9}{4}(1+\mu)(v_{c,0}^2 + v_{s,0}^2) - \frac{3}{2}\sqrt{3}(1-\mu)(u_{c,0}\, v_{c,0} + u_{s,0}\, v_{s,0}) \quad . \tag{34}$$

Now the difference of the Jacobi constants C_0 and C, of the periodic and non-periodic Trojan, follows from the respective Jacobi integrals as

$$C_0 - C = 2(\Omega_0 - \Omega) + (V^2 - V_0^2) \quad . \tag{35}$$

Since the right-hand side of this equation, consisting of constant as well as of periodic terms of many different periods, must always be equal to the constant left-hand side, it follows that all the variable terms with the same argument must separately add up to zero, and that the constant terms must also satisfy the relation

$$C_0 - C = 2(\overline{\Omega_0 - \Omega}) + \overline{(V^2 - V_0^2)} \quad . \tag{36}$$

To evaluate this expression for $\overline{C_0 - C}$, the results of Eqs. (32) and (34) have merely to be substituted. If one takes advantage of the relations existing between the coefficients of the elliptic fluctuation represented by Eqs. (28), and if, in line with the approximations already introduced, terms of order μ are omitted, the result from Eq. (36) is reduced to

$$C_0 - C \approx +\frac{16}{7}\,(v_{c,0}^2 + v_{s,0}^2)\ . \qquad (37)$$

This result confirms the numerical findings that the Jacobi constant C of the non-periodic orbit is always smaller than the Jacobi constant C_0 of the proper reference orbit, and that the difference C_0-C increases substantially with the amplitude of the short-period fluctuation as represented by $v_{c,0}$ and $v_{s,0}$. It should be noted that a relatively small change in C may be related to a rather large change in the amplitude of the predominant long-period libration. Since the right-hand side of Eq. (37) is roughly proportional to e^2, this relation is quite compatible with Eq. (1), according to which C is also a function of e^2.

The constants K and λ involved in Eqs. (30) can be approximated as follows: It can be seen that, with the approximations listed in Eqs. (18) for the third order partials of Ω, the constant contributions from the second order terms in Eqs. (9) amount to

$$K = -\frac{12}{7}\,(v_{c,0}^2 + v_{s,0}^2)\ , \quad \lambda = +\frac{12}{7}\sqrt{3}\,(v_{c,0}^2 - v_{s,0}^2)\ , \qquad (38)$$

neglecting again the higher order terms in u and v. Substitution of K and λ into the Eqs. (30) leads to the solution

$$u_{00} = +\frac{8}{7}\,(v_{c,0}^2 + v_{s,0}^2)\ , \quad v_{00} = -\frac{8}{7}\frac{1}{\sqrt{3}}\,(v_{c,0}^2 + v_{s,0}^2)\ . \qquad (39)$$

These small constant terms tend to produce an asymmetry of the principal short-period oscillation, just as the libration of long period has an asymmetry which increases

with the amplitude of the libration. The constant displacements u_{00}, v_{00} depend likewise on the (arbitrary) size of the principal short-period oscillation as represented by $v_{c,0}$ and $v_{s,0}$.

V. Summary and Conclusions.

It seems that the results of this analysis are of interest first insofar as they afford a deeper insight into the basic nature and stability of the non-periodic motions in a wide neighborhood of the triangular points. Secondly, however, it is hoped that the theory can be applied to the motions of the actual Trojan planets, after its proper extension to the less restricted case of an eccentric orbit of Jupiter, and then to the inclusion of deviations from the sun-Jupiter orbital plane. It will be appropriate, of course, to test the convergence of the expansions involved on a suitable hypothetical Trojan planet, with realistically chosen amplitude values for the librational and oscillatory motion. The principal advantage of this method undoubtedly lies in the ready availability of completely rigorous periodic reference orbits, which should eliminate at the outset a great deal of the work required in previous methods. The approach leads, of course, to a mixture of numerical and analytical features, which, however, should be no disadvantage, in the light of similarly constructed theories of other problems in celestial mechanics.

References

Rabe, E. 1961, Astron. J. 66, 500.
----. 1962, ibid. 67, 382.
Rabe, E. and Schanzle, A. 1962, ibid. 67, 732.

ELEMENTS OF A THEORY OF LIBRATIONAL MOTIONS IN THE ELLIPTICAL RESTRICTED PROBLEM

by

Professor E. Rabe
University of Cincinnati Observatory
Cincinnati, Ohio

I. Introduction.

In the preceeding lecture (Rabe, 1963), the non-periodic motions in the neighborhood of the periodic solutions of long period have been represented in the form of a series of superimposed oscillations, of various short and long periods, taking advantage of the availability of precise Fourier series representations for the long-periodic libration orbits in the restricted problem. If the relative motion of the two finite masses is assumed to be elliptic, instead of circular, these periodic solutions cease to exist. We know, however, that the two triangular points themselves remain solutions also in the elliptical problem. In this case, the equilateral configuration of the three masses remains the same at all times, and only their mutual distances experience periodic fluctuations of the order of the eccentricity e' of the relative orbit of the two finite masses. The three masses involved remain "at rest" in a non-uniformly rotating system of reference axes, if the system is conceived as pulsating, in addition to its non-uniform rotation. If the small particle is not exactly at one of the two equilateral points, but close to it, its motion relative to the two finite masses, therefore, may perhaps advantageously be described in terms of displacements from some appropriate "librational" orbit which participates in the non-uniform rotation and pulsation of such a system. It is attempted in this study to discuss the necessarily non-periodic motion of any small mass, with starting conditions comparable to those in the neighborhood of the periodic librations of the restricted problem, by first transposing the periodic orbits of the restricted problem into the

just described elliptical system, simply applying the periodic scale factor as determined by Jupiter's elliptic motion, and by then establishing the differential equations for the deviation of the true motion from this pulsating or modified intermediate orbit, which itself, of course, cannot be expected to be a solution of the differential equations of the elliptic problem. It may be anticipated, however, that the displacements can be found as a conglomerate of oscillations of various periods.

II. Differential Equations for Non-Uniformly Rotating Axes.

For convenience of reference, the two finite masses involved will be called sun and Jupiter, and the body of negligibly small mass a Trojan planet, but the derivations are valid, of course, also for the elliptical earth-moon problem (in the absence of the sun), or for any mass ratio permitting the existence of periodic solutions in the restricted case of the problem.

The center of our (ξ, η)-system of rectangular axes shall be assumed to coincide with the center of mass of sun and Jupiter, and the ξ-axis shall permanently coincide with the straight line connecting these two finite masses. Then, if

$$\rho = \frac{1-e^2}{1+e\cos\theta} \qquad (1)$$

represents the variable distance between sun and Jupiter as a function of the true anomaly θ and of the orbital eccentricity e, the unit of distance being identified with

the semi-major axis a=1 of Jupiter's heliocentric orbit, the coordinates (ξ_1, η_1) of the sun and (ξ_2, η_2) of Jupiter, in this non-uniformly rotating system, will be given by

$$\xi_1 = \frac{\mu}{1+\mu}\rho \qquad \eta_1 \equiv 0, \qquad \xi_2 = -\frac{1}{1+\mu}\rho \qquad \eta_2 \equiv 0, \quad (2)$$

where μ again denotes the mass of Jupiter in units of the solar mass. In some deviation from the introductory remarks, a fixed unit of distance has been defined, because the intended pulsation of the transposed reference libration will be achieved by the application of the factor ρ, as defined in Eq. (1).

As in the restricted problem, the Trojan's distance from sun and Jupiter will be denoted by Δ_1 and Δ_2, respectively, so that in terms of the Trojan's rectangular coordinates

$$\Delta_1^2 = (\xi - \xi_1)^2 + \eta^2$$
$$\Delta_2^2 = (\xi - \xi_2)^2 + \eta^2 \; . \qquad (3)$$

The unit of time shall be fixed again by demanding the constant of gravitation to be unity, so that Jupiter's mean motion N is given by

$$N = \sqrt{1+\mu} \; . \qquad (4)$$

In the non-uniformly rotating system just introduced, as associated with Jupiter's elliptic motion around the sun, the differential equations of motion for the Trojan are

$$\ddot{\xi} - 2\dot{\theta}\dot{\eta} - \dot{\theta}^2\xi - \ddot{\theta}\eta = \frac{\partial R}{\partial \xi}$$
$$\ddot{\eta} + 2\dot{\theta}\dot{\xi} - \dot{\theta}^2\eta + \ddot{\theta}\xi = \frac{\partial R}{\partial \eta}\,, \qquad (5)$$

if the motion is limited to Jupiter's orbital plane. The force function R in Eqs. (5) is given by

$$R = \frac{1}{\Delta_1} + \frac{\mu}{\Delta_2}\,. \qquad (6)$$

Now, since

$$\dot{\rho} = \rho^2 \frac{e \sin\theta}{1-e^2}\,\dot{\theta} \qquad (7)$$

and

$$\rho^2\dot{\theta} = N\sqrt{1-e^2}\,,$$

or

$$\dot{\theta} = \frac{N\sqrt{1-e^2}}{\rho^2}\,, \qquad (8)$$

one easily finds

$$\dot{\rho} = \frac{Ne}{\sqrt{1-e^2}} \sin\theta\,, \qquad (9)$$

$$\ddot{\rho} = \frac{N^2 e}{\rho^2} \cos\theta\,, \qquad (10)$$

$$\ddot{\theta} = -\frac{2N^2 e}{\rho^3} \sin\theta\,. \qquad (11)$$

In the corresponding, uniformly rotating (x,y)-system of the restricted problem sun-Jupiter-Trojan, let the periodic libration orbits again be represented in the form

$$x = x_0 + \sum_{j=1}^{\infty} x_{c,j} \cos(j\sigma) + \sum_{j=1}^{\infty} x_{s,j} \sin(j\sigma) ,$$
$$y = y_0 + \sum_{j=1}^{\infty} y_{c,j} \cos(j\sigma) + \sum_{j=1}^{\infty} y_{s,j} \sin(j\sigma) , \qquad (12)$$

with

$$\sigma = n(t - t_0) , \qquad n = \frac{2\pi}{T} , \qquad (13)$$

where T is the libration period and t_0 the zero epoch.

Since the motion to be transposed may be non-periodic already in the restricted problem, or, rather, a certain motion in the elliptical problem may be more closely related to a certain non-periodic trajectory in the restricted problem than to a periodic solution of the latter, the most reasonable transformation should have the form

$$\xi = \rho(x + u) , \qquad \eta = \rho(y + v) . \qquad (14)$$

Here x,y represent a periodic solution (12) of the restricted problem, ξ ,η are the coordinates of the elliptic problem as previously introduced, and u,v are the unknown deviations from the periodic solution x,y in the restricted problem, which have to be determined so that the ξ ,η as expressed in Eqs. (14) satisfy the differential equations (5).

It is convenient to introduce

$$\bar{x} = x + u , \qquad \bar{y} = y + v . \qquad (15)$$

In the restricted problem, these $\bar{x}$,$\bar{y}$ would be the coordinates of a non-periodic Trojan planet, satisfying the

differential equations

$$\ddot{\bar{x}} - 2N\dot{\bar{y}} = \frac{\partial \bar{\Omega}}{\partial \bar{x}} ,$$
$$\ddot{\bar{y}} + 2N\dot{\bar{x}} = \frac{\partial \bar{\Omega}}{\partial \bar{y}} , \qquad (16)$$

with

$$\bar{\Omega} = \bar{R} + \frac{1}{2}\left[N^2(\bar{x}^2 + \bar{y}^2) + \frac{\mu}{1+\mu}\right] = \bar{R} + \frac{1}{2}\left(\bar{\Delta}_1^2 + \mu\bar{\Delta}_2^2\right) , \qquad (17)$$

where the bar over Δ_1, Δ_2, and R indicates that these are the functions given by Eqs. (3) and (6) when ξ is replaced by $\bar{x}$, and η by $\bar{y}$.

On the basis of Eqs. (14) and (15), the time derivatives of ξ, η can be written in the form

$$\dot{\xi} = \dot{\rho}\bar{x} + \rho\dot{\bar{x}} \qquad (18)$$
$$\dot{\eta} = \dot{\rho}\bar{y} + \rho\dot{\bar{y}} ,$$
$$\ddot{\xi} = \ddot{\rho}\bar{x} + 2\dot{\rho}\dot{\bar{x}} + \rho\ddot{\bar{x}} \qquad (19)$$
$$\ddot{\eta} = \ddot{\rho}\bar{y} + 2\dot{\rho}\dot{\bar{y}} + \rho\ddot{\bar{y}}$$

In the Eqs. (3) for Δ_1 and Δ_2, not only ξ and η, but also ξ_1 and ξ_2 incorporate the variable factor ρ, so that

$$\Delta_1 = \rho\bar{\Delta}_1 , \qquad \Delta_2 = \rho\bar{\Delta}_2 , \qquad (20)$$

which relations are simply a consequence of the basic Eqs. (14), or

$$\xi = \rho\bar{x} , \qquad \eta = \rho\bar{y} , \qquad (21)$$

which introduce the effect of Jupiter's variable solar distance ρ on all the relative distances in the (ξ,η)-system. It is easily seen now that

$$\frac{\partial R}{\partial \xi} = \frac{1}{\rho^2}\frac{\partial \bar{R}}{\partial \bar{x}}, \qquad \frac{\partial R}{\partial \eta} = \frac{1}{\rho^2}\frac{\partial \bar{R}}{\partial \bar{y}}, \tag{22}$$

and if one substitutes Eqs. (18), (19), (22), as well as the earlier Eqs. (8) and (11) into the differential equations (5), considering also Eqs. (1) and (17), then, after multiplication of the resulting equations by ρ^2, these take the form

$$\begin{aligned}
&\rho^3\ddot{\bar{x}} - 2N\sqrt{1-e^2}\,\rho\dot{\bar{y}} + \frac{2N}{\sqrt{1-e^2}}\,\rho^2 e \sin\theta\dot{\bar{x}} = \frac{\partial\bar{\Omega}}{\partial\bar{x}} \\
&\rho^3\ddot{\bar{y}} + 2N\sqrt{1-e^2}\,\rho\dot{\bar{x}} + \frac{2N}{\sqrt{1-e^2}}\,\rho^2 e \sin\theta\dot{\bar{y}} = \frac{\partial\bar{\Omega}}{\partial\bar{y}}\,.
\end{aligned} \tag{23}$$

If these differential equations are compared with the earlier system of equations (16), with identical right-hand terms, the more involved nature of the left-hand sides of Eqs. (23) is due to the fact that Eqs. (23) are those for a u,v-solution satisfying the original Eqs. (5) of the elliptic problem, while Eqs. (16) determine the corresponding solution of the ordinary restricted problem. In order to obtain the differential equations for u and v, the $\bar{x},\bar{y}$ should now be separated into x,y, representing the periodic libration, and the increments u,v, according to Eqs. (15). At the same time, the following elliptic expansions shall be introduced into Eqs. (23), for all the periodic functions and constants depending on Jupiter's orbital motion:

$$\rho = 1 + \frac{1}{2}e^2 - (e - \frac{3}{8}e^3)\cos M - \frac{1}{2}e^2\cos 2M - \frac{3}{8}e^3\cos 3M\ldots$$
$$\rho^2 = 1 + \frac{3}{2}e^2 - 2(e - \frac{1}{8}e^3)\cos M - \frac{1}{2}e^2\cos 2M - \frac{1}{4}e^3\cos 3M\ldots \quad (24)$$
$$\rho^3 = 1 + 3e^2 - 3(e + \frac{3}{8}e^3)\cos M + \frac{1}{8}e^3\cos 3M\ldots ,$$

$$\sqrt{1-e^2} = 1 - \frac{1}{2}e^2\ldots$$
$$\frac{e\sin\theta}{\sqrt{1-e^2}} = (e - \frac{3}{8}e^3)\sin M + e^2\sin 2M + \frac{9}{8}e^3\sin 3M\ldots . \quad (25)$$

These expansions, in terms of the mean anomaly M of Jupiter in its orbit, are complete to the third order of e in the coefficients. M may be introduced as a linear function of time through

$$M = M_0 + N(t - t_0) , \quad (26)$$

where the epoch t_0 is identical with that used in Eqs. (12) and (13) for the periodic solutions x,y. Consequently, M_0 is the value of Jupiter's mean anomaly at an instant at which the periodic Trojan intersects the straight line connecting the equilateral point L_5 with the sun S on the outside of L_5. A the time t_0+T, when the periodic Trojan returns to this position, Jupiter's mean anomaly will in general have a value different from M_0, unless the values of N and n, or of the periods T and P, are commensurable like 13:1, 14:1, etc.

The right-hand sides of Eqs. (23) can be expanded as follows, about their corresponding periodic expressions

for u=v=0:

$$\frac{\partial \bar{\Omega}}{\partial \bar{x}} = \Omega_x + \Omega_{xx}u + \Omega_{xy}v + \tfrac{1}{2}\Omega_{xxx}u^2 + \tfrac{1}{2}\Omega_{xyy}v^2 + \Omega_{xxy}uv + \ldots$$
$$\frac{\partial \bar{\Omega}}{\partial \bar{y}} = \Omega_y + \Omega_{xy}u + \Omega_{yy}v + \tfrac{1}{2}\Omega_{xxy}u^2 + \tfrac{1}{2}\Omega_{yyy}v^2 + \Omega_{xyy}uv + \ldots . \qquad (27)$$

Here the notations $\Omega_x, \Omega_y, \Omega_{xx}, \ldots$ have been adopted for the corresponding partials with respect to x and y of the function Ω of the periodic Trojan of the restricted problem, and all these partials are periodic functions of time, through σ as given in Eq. (13).

Now the periodic solution x,y, on which all the partials of Ω in Eqs. (27) depend, satisfies the differential equations

$$\ddot{x} - 2N\dot{y} = \Omega_x ,$$
$$\ddot{y} + 2N\dot{x} = \Omega_y . \qquad (28)$$

After the substitutions and expansions have been made as described, and after the right-hand sides of Eqs. (23) have been expanded according to Eqs. (27), then the Eqs. (28) may be subtracted, and the following two differential equations result for u and v:

$$\ddot{u} - 2N\dot{v} = R_1 + e\left[E_1 + F_1\right]$$
$$\ddot{v} + 2N\dot{u} = R_2 + e\left[E_2 + F_2\right] . \qquad (29)$$

Here the right-hand sides have been divided into two principal parts, with a further subdivision of the eccentricity-affected second part into those terms independent of u,v,

and those depending on u, v. The three classes of terms so distinguished are represented by the following expressions:

$$R_1 = \Omega_{xx}u + \Omega_{xy}v + \tfrac{1}{2}\Omega_{xxx}u^2 + \tfrac{1}{2}\Omega_{xyy}v^2 + \Omega_{xxy}uv + \dots$$
$$R_2 = \Omega_{xy}u + \Omega_{yy}v + \tfrac{1}{2}\Omega_{xxy}u^2 + \tfrac{1}{2}\Omega_{yyy}v^2 + \Omega_{xyy}uv + \dots , \qquad (30)$$

$$E_1 = f\dot{u} + g\dot{v} + h\ddot{u}$$
$$E_2 = -g\dot{u} + f\dot{v} + h\ddot{v} , \qquad (31)$$

$$F_1 = f\dot{x} + g\dot{y} + h\ddot{x}$$
$$F_2 = -g\dot{x} + f\dot{y} + h\ddot{y} , \qquad (32)$$

with

$$f = -2N\left[(1 + \tfrac{3}{8}e^2)\sin M - \tfrac{1}{8}e^2 \sin 3M \dots\right]$$
$$g = -2N\left[(1 - \tfrac{7}{8}e^2)\cos M + \tfrac{1}{2}e\cos 2M + \tfrac{3}{8}e^2\cos 3M \dots\right] \qquad (33)$$
$$h = -3e + 3(1 + \tfrac{3}{8}e^2)\cos M - \tfrac{1}{8}e^2\cos 3M \dots .$$

It is seen that the Eqs. (30) for R_1 and R_2 are identical with the entire right-hand sides of the differential equations in the case of the restricted problem, and indeed Eqs. (29) reduce to those of the restricted problem when e=0. For small eccentricities, as that of Jupiter's orbit, the restricted problem (solution) for the motion of a non-periodic Trojan should therefore still represent a good first approximation to that part of the complete solution of the Eqs. (29) which may be called the free oscillation, as based on arbitrarily chosen starting data. Evidently the particular solution

$$u \equiv 0 , \qquad v \equiv 0 , \qquad (34)$$

representing the periodic libration x,y in the restricted case where e=0, does not exist when e≠0. As far as the "elliptic" terms represented by E_1 and F_1 are concerned, they clearly will produce additions of the order of e to the restricted problem solution, and these additions have to satisfy the complete Eqs. (29), including the parts denoted by R_1 and R_2.

The parts involving E_1 and E_2 depend on the first and second time derivatives of u and v, and therefore, once any solution including the consideration of these terms has been obtained, they have to be considered also in any subsequent approximations leading to terms of higher order. The parts involving F_1 and F_2, on the other hand, do not depend on u,v and their derivatives, but only on the short-period orbital motion of Jupiter and the long-period libration represented by the periodic solution (12). To find the effect of these terms, which cause a "forced" oscillation about the reference orbit, the R-parts of the differential equations have to be satisfied, of course, in the first approximation neglecting second and higher powers of the eccentricity e, while the consideration of the E-terms may and even should be postponed until the subsequent second approximation, because the terms produced by the substitution of the first approximation, of $O(e)$, into E_1 and E_2 are of $O(e^2)$. The determination of the forced oscillations, to any desired degree of perfection, may be achieved without any consideration of the additional free oscillations, and the successive approximations for the free oscillations can be obtained without any consideration of the F-terms, which are the source of the forced oscillations. The E-terms, however,

affect both parts of the solution, beginning with the initial consideration of their presence, and once included play a similar role as the R-terms.

In the light of these general considerations of the significance of the various parts in the differential equations (29), the general solution can be divided into two parts, in the form

$$u = u_0 + u_f \,, \qquad v = v_0 + v_f \,, \tag{35}$$

where u_0, v_0 represent the forced and u_f, v_f the free part of the solution. Now the earlier statement concerning the impossibility of permanently vanishing u,v can be qualified by saying that

$$u_0 \not\equiv 0 \,, \qquad v_0 \not\equiv 0 \tag{36}$$

constitutes the fundamental and fixed part of the solution, but that the particular case

$$u_f \equiv 0 \,, \qquad v_f \equiv 0 \tag{37}$$

is admissible for the free part of the solution. Evidently then the solution represented by u_0, v_0, even though non-periodic in nature, plays a role in the elliptic problem which is equivalent to the role of the periodic solution in the restricted problem.

III. Some Basic Features of the Solution.

As mentioned before, the first approximation to the

forced part of the solution can be obtained without the consideration of the E-terms, or from the reduced equations

$$\begin{aligned} \ddot{u}_0 - 2N\dot{v}_0 &= R_1 + eF_1 \ , \\ \ddot{v}_0 + 2N\dot{u}_0 &= R_2 + eF_2 \ . \end{aligned} \tag{38}$$

Since the right-hand sides of these equations involve products of Fourier series depending on multiples of σ with others depending on multiples of M, the solution must necessarily be of the form

$$\begin{aligned} u_0 &= \sum_j \sum_k u_{j,k} \exp\left[i(jM + k\sigma)\right] \\ v_0 &= \sum_j \sum_k v_{j,k} \exp\left[i(jM + k\sigma)\right] \ , \end{aligned} \tag{39}$$

where the integers j,k may have any value from $-\infty$ to ∞. The coefficients $u_{j,k}$, $v_{j,k}$ have to be determined from the identities which are the result of substituting Eqs. (39) into the differential equations (38). In contrast to the situation encountered in the determination of the coefficients of the principal terms of the free solution, which is identical to the situation in the first approximation for the solution in the restricted problem (Rabe, 1963), the identities resulting from the substitution into Eqs. (38) have absolute terms, produced by eF_1 and eF_2, respectively, and therefore the solution procedure will be rather straightforward. Successive approximations will be necessary, of course, but the convergence of the solution will benefit again from the rapid convergence of the series involved in R_1 and R_2, and, in the subsequent steps considering second and higher powers of e, from the convergence of the basic elliptic expansions (24) and (25).

For an illustration of the method for the determination of the coefficients in the solution (39), let it be assumed that the amplitude of the periodic libration x,y is small enough to justify the omission of all the periodic terms of Ω_{xx}, Ω_{yy}, and Ω_{xy} in the first approximation. Neglecting also in R_1 and R_2 all the terms involving the second and higher order powers and products of u_0 and v_0, as well as in the parts eF_1 and eF_2 of Eqs. (38) all the second and higher powers of e, these differential equations will be reduced to

$$\begin{aligned} \ddot{u}_0 - 2N\dot{v}_0 &= \alpha_0 u_0 + \gamma_0 v_0 - 2Ne\,(\dot{x}\sin M + \dot{y}\cos M)\,, \\ \ddot{v}_0 + 2N\dot{u}_0 &= \gamma_0 u_0 + \beta_0 v_0 + 2Ne\,(\dot{x}\cos M - \dot{y}\sin M)\,, \end{aligned} \tag{40}$$

where the α_0, β_0, γ_0 are the constant terms of Ω_{xx}, Ω_{yy}, Ω_{xy} (Rabe, 1963). It may be noted that the terms involving $\ddot{x}$ and $\ddot{y}$, which actually appear in F_1 and F_2 according to Eqs. (32), have been omitted, too, because they contain the second order factor n^2, as compared to the first order factor n contained in $\dot{x}$ and $\dot{y}$.

The Eqs. (40) are simple enough to assume the first approximation to the solution immediately in the trigonometric form

$$\begin{aligned} u_0 &= u_{c,1}\cos(M+\sigma) + u_{s,1}\sin(M+\sigma) + u_{c,-1}\cos(M-\sigma) + \\ &\quad u_{s,-1}\sin(M-\sigma) \\ v_0 &= v_{c,1}\cos(M+\sigma) + v_{s,1}\sin(M+\sigma) + v_{c,-1}\cos(M-\sigma) + \\ &\quad v_{s,-1}\sin(M-\sigma)\,, \end{aligned} \tag{41}$$

omitting terms with higher multiples of M and σ. It is

easily seen that 2M enters the solution only in connection with the e^2-terms in eF_1 and eF_2, and that 2σ enters only in connection with correspondingly smaller terms in $\dot{x}$ and $\dot{y}$ as well as in R_1 and R_2. Furthermore, terms independent of M or σ, or of both variables, do not appear in this first order approximation. Accordingly, the terms considered in Eqs. (41) clearly constitute the principal terms of the forced solution u_0, v_0.

The substitution into Eqs. (40) produces the following eight identities, conveniently divided into two groups of four equations each:

$$\begin{aligned}
A_1 u_{c,1} + C_1 v_{c,1} + D_1 v_{s,1} &= enN\,(x_{c,1} + y_{s,1}) \\
A_1 u_{s,1} - D_1 v_{c,1} + C_1 v_{s,1} &= enN\,(x_{s,1} - y_{c,1}) \\
C_1 u_{c,1} - D_1 u_{s,1} + B_1 v_{c,1} &= enN\,(y_{c,1} - x_{s,1}) \\
D_1 u_{c,1} + C_1 u_{s,1} + B_1 v_{s,1} &= enN\,(x_{c,1} + y_{s,1})
\end{aligned} \tag{42}$$

$$\begin{aligned}
A_{-1} u_{c,-1} + C_{-1} v_{c,-1} + D_{-1} v_{s,-1} &= enN\,(-x_{c,1} + y_{s,1}) \\
A_{-1} u_{s,-1} - D_{-1} v_{c,-1} + C_{-1} v_{s,-1} &= enN\,(x_{s,1} + y_{c,1}) \\
C_{-1} u_{c,-1} - D_{-1} u_{s,-1} + B_{-1} v_{c,-1} &= enN\,(-x_{s,1} - y_{c,1}) \\
D_{-1} u_{c,-1} + C_{-1} u_{s,-1} + B_{-1} v_{s,-1} &= enN\,(y_{s,1} - x_{c,1}) \,.
\end{aligned} \tag{43}$$

The coefficients denoted by A_1, B_1, C_1, D_1, A_{-1}, B_{-1}, C_{-1}, D_{-1}, are as follows:

$$\begin{aligned}
A_1 &= \alpha_0 + (N+n)^2 & A_{-1} &= \alpha_0 + (N-n)^2 \\
B_1 &= \beta_0 + (N+n)^2 & B_{-1} &= \beta_0 + (N-n)^2 \\
C_1 &= C_{-1} = \gamma_0 & & \\
D_1 &= 2N\,(N+n) & D_{-1} &= 2N\,(N-n)\,.
\end{aligned} \tag{44}$$

The coefficients of the unknowns on the left-hand sides of the Eqs. (43) differ from those of the unknowns on the left-hand sides of Eqs. (42) only by the appearance of (N-n) instead of (N+n). The right-hand sides of both sets involve different combinations of the principal coefficients of $\dot{x}$ and $\dot{y}$, and consequently the factor n, in addition to the presence of e (and N) as factors. In order to have unique solutions of the two sets of linear equations, for the various unknown coefficients of the Eqs. (41), the determinants of the A_1, B_1, C_1, D_1, and of the A_{-1}, B_{-1}, C_{-1}, D_{-1}, should not vanish. Denoting the first determinant by π_1, the second one by π_{-1}, the resulting expressions can be represented in one equation:

$$\pi_{1,-1} = \left\{\left[\alpha_0 + (N\mp n)^2\right]\left[\beta_0 + (N\mp n)^2\right] - \gamma_0^2 - 4N^2(N\mp n)^2\right\}^2 . \tag{45}$$

With the approximating values of α_0, β_0, γ_0 valid at the libration point L_5 (Rabe, 1963), the contents of the large $\{\ \}$ bracket in Eq. (45) are found to approximate $\mp 2n$, and all the not-vanishing sub-determinants of the first order are then also approximated by 2n or -2n. Consequently, with

$$\pi_{1,-1} \approx 4n^2 , \tag{46}$$

a small divisor of the order of n affects the solution of the two linear systems (42) and (43), at least for the relatively small libration amplitudes where α_0, β_0, γ_0 can be approximated by the values of Ω_{xx}, Ω_{yy}, Ω_{xy} at L_5. The coefficients $u_{c,1}$, $u_{s,1}$, ..., $v_{s,-1}$ of the solution (41) are of the order of e·L, where L represents the amplitude of the basic libration of long period.

To find the second approximation to the forced solution u_0, v_0, including coefficients involving e^2; $x_{c,2}$, ..., $y_{s,2}$; α_1, α_{-1}, ..., γ_{-1}, the result of the first approximation (41) will have to be substituted into the various previously neglected terms of the differential equations (29). Careful planning will be necessary to make sure that, depending on the numerical values of the various basic coefficients involved, the proper coefficients, products and powers of small quantities are considered in each successive approximation. It can be seen, however, that the substitution of the first approximation into the previously neglected terms of higher order will create new terms, of order e^2 etc., which in general will have arguments of the form $(jM + k\sigma)$ considered in Eqs. (39), but including terms where either j or k, or both, may be zero. Consequently, certain terms of long periods will emerge with the terms of order $e^2 \cdot L$. The constant terms produced by the substitution can be absorbed by corresponding small constant terms in u_0, v_0, just as in the case of the second approximation to the free solution in the restricted problem (Rabe, 1963). In general, the previous approximation to the forced solution will have to be adjusted for any terms with the same arguments obtained from any subsequent additive solution. While the details of the successive approximations will depend on the amplitude of the basic libration orbit, the present exploratory analysis indicates the stable nature of these forced oscillations, which in their entirety determine one unique and particular orbit in the elliptic problem, namely the equivalent of the selected periodic libration of the restricted problem.

In the general solution of the complete Eqs. (29),

the free part of the solution has to be added and can be evaluated for any reasonable starting conditions. The principal part of the free solution, as has been seen, is identical with the corresponding approximation in the restricted problem, where the forced part of the solution is non-existent. The second approximation to the free solution in the elliptical problem should, as has also been shown, include consideration of the E-terms in Eqs. (29). Since these E-terms involve periodic functions of M (through f,g,h), which are multiplied by the derivatives of u and v as obtained from the first approximation, and since the principal terms of the first approximation are periodic functions of $c\sigma$, the second order terms now created on the right-hand sides of Eqs. (29) will involve arguments of the forms

$$M + c\sigma , \qquad M - c\sigma , \text{ etc.}$$

The periods of M and $c\sigma$ differ only by amounts of the order of μ. More precisely, one has

$$N = \sqrt{1+\mu} , \qquad cn \approx 1 - \frac{23}{4}\mu , \tag{47}$$

the second of these relations being an approximation. To the same degree of accuracy one gets therefore

$$M - c\sigma = M_0 + \frac{25}{4}\mu\ (t - t_0) . \tag{48}$$

Those terms of the free solution u_f, v_f depending on the argument $(M-c\sigma)$ will have the extremely long period T* determined by the (approximate) frequency

$$\nu^* = \frac{25}{4}\mu . \tag{49}$$

Consequently,

$$T^* = \frac{8\pi}{25\mu} , \qquad (50)$$

which amounts to more than 12 libration periods. The effect of such terms will be small, however, since they appear only in the "elliptical" second-order part of the free solution. Their actual size will depend, as the amplitudes of all the free oscillations, on the initial displacement from the forced solution, and consequently approaches zero when the initial deviations are reduced to zero.

Reference

Rabe, E. 1963, "Outline of a Theory of Non-Periodic Motions in the Neighborhood of the Long-Period Librations about the Equilateral Points of the Restricted Problem of Three Bodies". Lecture, delivered at the Summer Seminar in Space Mathematics, Cornell University (July 29, 1963).

SHOCK WAVES IN RAREFIED GASES

by

S. F. Shen

Cornell University

I. Introduction

In this series of lectures we shall survey the developments in rarified gasdynamics toward the solution of flow problems, the shock wave structure serving as an example illustrating the difficulties that led to the various refinements and alternatives. By rarified gasdynamics is meant the branch of gasdynamics which cannot be dealt with by the conventional continuum theory of a viscous and heat-conducting gas, hereafter referred to as simply the (conventional) "continuum theory", because of the effects of very low density. The concept of a continuum however is usually still adopted, but modifications are required in two main aspects: Firstly, the law relating the viscous stresses to fluid deformation (the Navier-Stokes relations) and that relating the heat flux to temperature gradient (the Fourier law) are theoretically no longer valid. Secondly, the boundary conditions of "no velocity slip" and "no temperature jump" at a solid boundary, generally assumed in conventional continuum theory, must be re-examined. By restricting ourselves to the problem of the shock wave structure, we essentially divorce ourselves from the latter question. Our efforts therefore are directed toward only a formulation of the proper equations to be used in rarified gasdynamics.

To seek a logical theory which is capable of treating the departure from the conventional Navier-Stokes and Fourier laws due to the very low density, we fall back on the kinetic theory of gases. The gas is now regarded as consisting of numerous molecules interacting with each other and with the environment according to the laws of classical mechanics. For the simplest case of a monatomic gas, the molecules are all alike and have spherical symmetry. This will be understood as our model in the following discussion. The phenomenal success of kinetic theory in predicting quantitatively, with suitable chosen force laws between molecules, the viscosity and heat conduction coefficients for use in conventional continuum flows is well-known. Equally confirmed are its deductions concerning the flow in the "free molecule" limit of very, very low densities, such as regarding such flows through orifices or capillaries. These being the two extremes of the spectrum, we expect that it should also be fruitful in intermediate stages that characterize rarified gasdynamics.

II. Flow Regimes and the Knudsen Number

When we consider a body of gas enclosed in a vessel of volume V, in equilibrium, the state of the gas is defined thermodynamically by the pressure p, the density ρ, and/or temperature T. These quantities must first be defined from the viewpoint of kinetic theory. Let each of the molecules have a mass m, and the total number in V be N. The density follows directly as

$$\rho = Nm/V = nm$$

where $n \equiv N/V$, the number density. If each molecule is characterized by its "size" σ, which may be the diameter for the simplest model of hard sphere molecules, the average spacing λ' of the N molecules occupying volume V is

$$N\lambda'^3 \sim V$$

or $\lambda' \sim n^{1/3}$. We have thus a dimensionless parameter for the degree of rarefaction of the gas as λ'/σ, i.e., the average size of the cell for each molecule in terms of the molecular diameter. In classical kinetic theory, this parameter is shown to be related directly in the corrections of the perfect gas law:

$$p = \rho RT \ .$$

We shall, however, assume that λ'/σ is sufficiently large that the perfect gas law holds. A typical value of σ is 10^{-8} cm. At standard conditions (0°C and 1 atm.), the number density of gas molecules is given by the Loschmidt number,

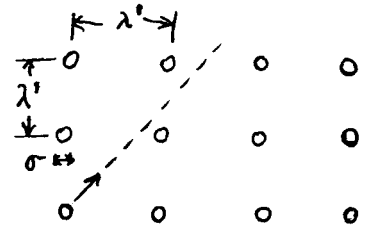
$$h = 2.69 \times 10^{19} / \text{cm.}^3$$

or

$$\lambda' \cong \frac{1}{3} \times 10^{-6} \text{ cm.}$$

and $\lambda'/\sigma \cong 30$. The ratio becomes obviously larger as the density is decreased, since σ remains unchanged.

The parameter λ'/σ is "static" in nature. As molecules are actually continuously in motion, there is a "dynamic" characteristic length representing the average distance travelled by a molecule between successive collisions, known as the "mean free path" λ. Imagine the molecules arranged actually at distance λ' apart in a regular pattern. When a



molecule moves in an arbitrary direction, the probability of its hitting a second molecule at a distance of $O(\lambda')$ is proportional to the ratio of the target area σ^2 to the passage area λ'^2. Hence we expect

$$\lambda \sim \lambda' O(\lambda'^2/\sigma^2) \sim O(1/n\sigma^2).$$

With the same typical values for λ' and σ the estimate at 0°C and 1 atm. is

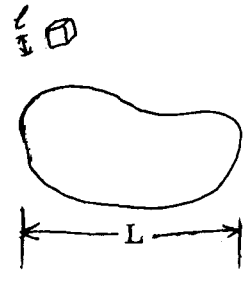
$$\lambda \sim (2.69 \times 10^{19})^{-1} \times 10^{16} = 10^{-3} \text{ cm.}$$

The mean free path goes up quickly as the density is reduced. In the standard atmosphere at 100 miles altitude, for instance,

$$\lambda' \cong 10^{-4} \text{ cm.}$$

$$\lambda \cong 300 \text{ cm.}$$

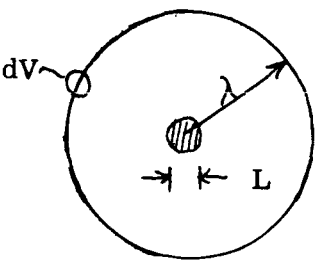
Now gasdynamics deals always with flow problems, involving therefore in addition a characteristic length which represents the scale of the flow phenomenon. To fix our ideas, let us imagine the flow as over a body of length L. We assume the flow can be described mathematically as a continuum, so that it must be possible to introduce "fluid elements" of size ℓ with $\ell \ll L$. On the other hand, to apply kinetic theory, it is necessary that statistical properties over molecules are well defined in a fluid element. In other words, there must be a sufficient number of molecules in a cell of size ℓ, or $\ell \gg \lambda'$. Thus rarefied gasdynamics generally deals with the restriction of $L \gg \ell \gg \lambda' \gg \sigma$. Because of the large difference of orders of magnitude between λ and λ', the value of λ may be still taken as arbitrary.



The significance of the mean free path is that it is a measure of the memory of the individual molecules in a flow field of size L. When a molecule hits the body, it eventually rebounds after taking on some characteristics peculiar to the body surface. The explicit dependence of the molecular motion on the body characteristics is therefore confined in a "sheath" of thickness roughly $O(\lambda)$ surrounding the body, which may be referred to as the "Knudsen layer". Beyond the Knudsen layer the body influence is only indirect, being propagated through successive collisions of the molecules which never were in direct contact with the body. The conventional continuum

equations of motion display no explicit dependence on the body geometry and its properties. It seems clear that they at most are applicable to the region beyond the Knudsen layer. If boundary conditions are nevertheless stipulated at the body, the implication must be a vanishingly thin Knudsen layer. The case consequently corresponds to the limit of $\lambda/L \to 0$. The parameter λ/L is known as the Knudsen number Kn.

In the other limit of $Kn \to \infty$, the Knudsen layer extends to a sphere of radius λ, the body being shrunk to a small region of size L near the origin. Of all molecules crossing the spherical surface of area $O(\lambda^2)$, only a small fraction $O(L^2/\lambda^2)$ has collided with the body and rebounded to cross the sphere again. Hence, if we examine the composition of the molecules in any small volume element dV in the neighborhood of the spherical surface, there is hardly any that comes directly from the body. The flow in the region outside of the sphere is described by conventional continuum theory, but it is now almost undisturbed by the presence of the body. Such considerations lead to the "free molecule flow" approximation, where momentum and energy transfers to the body are evaluated as if the free stream were completely undisturbed by collisions caused by those molecules rebounding from the body. It may be noted, however, that the "free molecule flow" approximation is never valid for the flow field at distances away from the body large compared to the mean free path. In particular, for two-dimensional motions in the x, y-plane, say, many molecules frim the z-direction certainly have suffered collisions before arriving at the plane of motion. This has led to difficulties, for instance, in the free molecule flow through a two-dimensional channel.

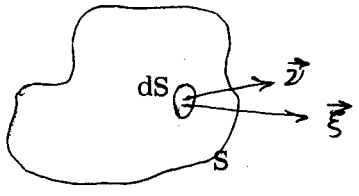


In between the limits of continuum and free-molecule flows, the flow regimes are often classified according to the magnitude of the Knudsen number following Tsien (1946). Thus for $Kn \ll 1$ the flow is said to be in the "slip flow" regime, in the sense that only the "no slip" and "no jump" conditions at a solid boundary need to be modified, but the equations of motion remain unchanged. Beyond the "slip flow" regime and before reaching the flow molecule limit, i.e., for Kn $O(1)$, the flow is said to go through a "transition regime". It is in fact for flow problems in this regime that

much work remains to be done.

III. Kinetic definition of pressure and temperature of gas in equilibrium. The Maxwellian distribution function

We return to the question of defining pressure and temperature for gas in equilibrium in terms of the motion of the molecules. The gas is assumed to be in a fixed vessel, and the pressure is uniform. In the interior consider a small volume element dV enclosed by control surface S. Through a small area dS on the surface, molecules having velocity $\vec{\xi}$ will pass at the rate of $n_{\vec{\xi}}\, \vec{\xi} \cdot \vec{\nu}\, dS$, where $n_{\vec{\xi}}$ is the number density of such molecules, and $\vec{\nu}$ the unit outward normal on dS. The rate of momentum loss due to such molecules in the $\vec{\nu}$ direction is therefore $m n_{\vec{\xi}} (\vec{\xi} \cdot \vec{\nu})^2\, dS$. For all molecules having various velocities $\vec{\xi}$, the total rate of momentum loss is obtained by summing over all $\vec{\xi}$'s. The result on the other hand must be equivalent to the action of a pressure force $p\,dS$ on the same surface element. Hence we write



$$p = \sum_{\vec{\xi}} n_{\vec{\xi}}\, m\, (\vec{\xi} \cdot \vec{\nu})^2 .$$

For a comparable definition of the thermodynamic temperature, we consider again the gas in a fixed vessel of volume V. Let us imagine heat has been added to raise the temperature from absolute zero to T. Kinetically, all this energy, say E, can only go into the translation energy of the monatomic molecules. Thus

$$E = \int_0^T C_v\, dT \cdot \rho V = \sum_{\vec{\xi}} n_{\vec{\xi}}\, V \cdot \frac{1}{2} m \xi^2$$

or

$$\rho \int_0^T C_v\, dT = \sum_{\vec{\xi}} \frac{1}{2} m n_{\vec{\xi}}\, \xi^2$$

where C_v is the specific heat at constant volume, per unit mass, and generally temperature dependent. This gives an implicit definition of T in terms of the molecular motion.

Since it is always assumed that a large number of molecules are present in

the small control volume dV, we must regard $\vec{\xi}$ as covering the entire range between zero and infinity. The summation over $\vec{\xi}$ therefore goes over into an integration. The number density $n_{\vec{\xi}}$ may be rewritten as

$$n_{\vec{\xi}} = n f(\vec{\xi})\, d\vec{\xi} ,$$

n being the number density of all molecules, and $f(\vec{\xi})\, d\vec{\xi}$ giving the fraction having velocity between $\vec{\xi}$ and $\vec{\xi} + d\vec{\xi}$. The symbol $d\vec{\xi}$ above should be understood as a volume element in the velocity space, e.g., in Cartesian space (ξ_1, ξ_2, ξ_3):

$$d\vec{\xi} = d\xi_1\, d\xi_2\, d\xi_3$$

and not a differential vector. The function $f(\vec{\xi})$ is referred to as the "velocity distribution function".

Obviously, since

$$n = \sum_{\vec{\xi}} n_{\vec{\xi}} = \int_{\vec{\xi}} n f(\vec{\xi})\, d\vec{\xi} = n \int_{\vec{\xi}} f(\vec{\xi})\, d\vec{\xi}$$

the velocity distribution function has the property

$$\int_{\vec{\xi}} f(\vec{\xi})\, d\vec{\xi} = 1 . \qquad \text{(III.1)}$$

Turning to the pressure definition, we have

$$\begin{aligned} p &= \int_{\vec{\xi}} n f d\vec{\xi}\, m (\vec{\xi} \cdot \vec{\nu})^2 \\ &= \rho \langle \xi_\perp^2 \rangle \end{aligned} \qquad \text{(III.2)}$$

where $\xi_\perp \equiv \vec{\xi} \cdot \vec{\nu}$, the velocity component normal to the surface, and the bracketed quantity $\langle Q \rangle$ represents in general the average of the property Q over all molecules in the velocity space,

$$\langle Q \rangle \equiv \int_{\vec{\xi}} Q f(\vec{\xi})\, d\vec{\xi} . \qquad \text{(III.3)}$$

Likewise, for the temperature we get

$$\rho \int_0^T C_v\, dT = \frac{\rho}{2} \langle \xi^2 \rangle . \qquad \text{(III.4)}$$

It is next of interest to deduce the velocity distribution function in a gas in equilibrium. First of all there can be obviously no dependence on orientation, so the distribution function must be a function of the speed ξ alone. Let the entire speed

range be divided into a finite number of discrete cells according to the average $\xi^{(i)}$ within the cell, and let the number of molecules within the ith cell be a_i. It is then postulated that at equilibrium the distribution of the molecules is the most probable random arrangement of the N molecules into all these cells, subject to the constraints that the total number N and the total kinetic energy E be kept constant.

If we assign a set of numbers a_i, the number of ways to achieve such an arrangement is $N!/\prod_i (a_i!)$. Therefore we write the possibility P of such an arrangement as $P = N!/\prod_i (a_i)!$ For convenience, the maximization may be carried out for $P' = \log P$ under the constraints $N = \text{const.}$ and $E = \text{const.}$ Using Lagrange's multipliers α and α', we finally seek to maximize $P' - \alpha N - \alpha' E$. The necessary condition is thus

$$\frac{\partial}{\partial a_i} P' - \alpha - \alpha' \cdot \frac{1}{2} m \xi^{(i)2} = 0 .$$

But $$\frac{\partial}{\partial a_i} P' = -\frac{\partial}{\partial a_i} \log (a_i)! \text{ , the "logarithmic derivative",}$$

$$\cong -\left[\log a_i + \frac{1}{2a_i} + \cdots \right]$$

for $a_i >> 1$ (Jahnke and Emde: "Tables of Functions", p. 18). Hence assuming $a_i >> 1$,

$$\log a_i = -\alpha - \frac{1}{2} \alpha' m \xi^{(i)2}$$

or $a_i = A \exp\left(-\frac{\alpha'}{2} m \xi^{(i)2} \right)$. If the cell size is made to tend to zero formally, the distribution functions must be of the form

$$f = A e^{-\beta \xi^2} \qquad \text{(III.5)}$$

known as the Maxwellian distribution function. The two constants A and β are determined by the constants N and E. To integrate, note that the volume element $d\vec{\xi}$ should be evaluated as that of the spherical shell between speeds ξ and $\xi + d\xi$, i.e., $4\pi \xi^2 d\xi$. In this way we find

$$A = (\beta/\pi)^{3/2}, \quad \frac{3}{4\beta} = E = \int_0^T C_v dT . \qquad \text{(III.6)}$$

Meanwhile, for the pressure p, we choose $\vec{v}$ to be in the Cartesian x-direction and denote the velocity $\vec{\xi}$ by its Cartesian components (ξ_1, ξ_2, ξ_3). Then

$$p = \rho \langle \xi_\perp^2 \rangle = \rho A \int_0^\infty\int\int \xi_1^2 e^{-\beta(\xi_1^2 + \xi_2^2 + \xi_3^2)} d\xi_1 d\xi_2 d\xi_3 = \rho/2\beta \ .$$

Since the perfect gas law is assumed to hold, the pressure formula gives

$$\beta = 1/2\,RT \ . \qquad \text{(III.7)}$$

From the second of Eq. (III.6)

$$C_v = \frac{\partial E}{\partial T} = \frac{\partial}{\partial T}(3/4\beta) = \frac{3}{2}R \ ,$$

a well-known result in thermodynamics.

It should be noted that the assumption of $a_i >> 1$ that led to the Maxwellian distribution is clearly violated as $\xi^{(i)} \to \infty$. In fact, although integrations in the velocity space are always formally carried out to $\xi \to \infty$, the logical cut-off for a given E cannot exceed $\xi_{max.} = \sqrt{2E/m}$, which is the speed of a single molecule absorbing the entire amount of energy. The assumption of $a_i >> 1$ ceases to hold before $\xi^{(i)} = \xi_{max.}$, and the Maxwellian distribution function has little significance for molecules whose speeds approach $\xi_{max.}$. It however applies to <u>almost</u> all molecules.

With the Maxwellian distribution function, the state of a gas is fully described by the two parameters n and β. Instead of β, it is often more physical to use the average speed of the molecules $\bar{c}$,

$$\bar{c} \equiv \langle \xi \rangle = \sqrt{8RT/\pi} \qquad \text{(III.8)}$$

which is quite close to the sound speed a ($a = \sqrt{\gamma RT}$, $\gamma = 5/3$ for monatomic gases), which is the propagation speed of small disturbances and plays an important role in conventional gasdynamics. Likewise the number density n is directly related to the mean free path λ. The state of the gas molecules thus may also be characterized by $\bar{c}$ and λ. Out of $\bar{c}$ and λ, we can further construct a time constant τ or its inverse Θ:

$$\tau = \lambda/\bar{c} = 1/\Theta. \qquad \text{(III.9)}$$

Θ is known as the "collision frequency" of a molecule, since in time δt, $\Theta\delta t$ gives the average distance travelled $\bar{c}\delta t$ divided by the average distance λ between collisions.

The time constant τ is of considerable interest. If the gas is not in equilibrium, it is plausible to imagine that collisions tend to bring the gas to the most

probable, hence the equilibrium distribution. The time constant for this process is no other than τ. In the case of gas in non-uniform motion, there is also a time constant for the overall phenomenon. If the latter is much greater than τ, at each instant and location the molecules in a small volume element dV will be in "quasi-equilibrium". That is, as a first approximation, the velocity distribution should be Maxwellian, with n and β assuming the instantaneous local values, but the observer must now ride with the average velocity U (over all the molecules in dV). We denote this as the "local Maxwellian" $f^{(0)}$,

$$f^{(0)} = \left(\frac{\beta}{\pi}\right)^{3/2} e^{-\beta[\vec{\xi} - \vec{U}]^2} = \left(\frac{\beta}{\pi}\right)^{3/2} e^{-\beta \vec{c}^2} \qquad \text{(III.10)}$$

where $\vec{c} \equiv \vec{\xi} - \vec{U}$, sometimes referred to as the "thermal velocity". It is easy to verify that following Eq. (III.10), $\langle \vec{\xi} \rangle = \vec{U}$ as stated; also $\langle \vec{c}^2 \rangle = 3/2\beta$, showing that β (or T) is intimately connected with $\vec{c}$.

As a further illustration, in a slightly non-uniform gas let us assume that the state of the molecular motion be still characterized approximately by $\bar{c}$ and λ. Together with the molecular properties of mass m and size σ, there are now four parameters from which, among other things, the behavior of the transport properties may be deduced at least qualitatively. For the viscosity coefficient μ, suppose

$$\mu = \mu(m, \sigma, \bar{c}, \lambda) .$$

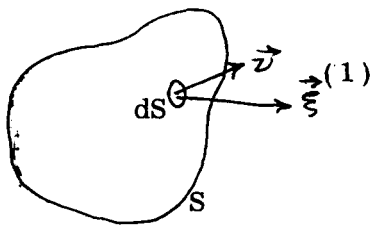
By dimensional reasoning, there must be

$$\mu \sim \frac{m\bar{c}}{\lambda^2} F(\lambda/\sigma) \sim \rho \lambda \bar{c}\, G(\lambda/\sigma) .$$

The function $G(\lambda/\sigma)$ should be taken in the limit $\lambda/\sigma \to \infty$. Thus the first approximation should yield $\mu \sim \rho \bar{c} \lambda$, which is confirmed by detailed analysis.

IV. The Boltzmann Equation

The statement that the velocity distribution function for a gas not in equilibrium is subject to change due to molecular collisions is mathematically expressed by the Boltzmann equation. Consider an arbitrary control volume V enclosed by the surface S in the interior of the gas. On a surface element dS let $\vec{\nu}$ be a unit



outward normal. For those molecules having velocity between $\vec{\xi}^{(1)}$ and $\vec{\xi}^{(1)} + d\vec{\xi}^{(1)}$, the total number in V at any time is $\int_V n f(\vec{\xi}^{(1)})\, d\vec{\xi}^{(1)}\, dV$. The flux of such molecules through the surface S is $\int_S (\vec{\xi}^{(1)} \cdot \vec{\nu})\, n f(\vec{\xi}^{(1)})\, d\vec{\xi}^{(1)}\, dS$. Let further the rate that such molecules are created in a small volume element dV, through collisions, be denoted by $\frac{\partial}{\partial t} n f_1 \Big|_{coll.} d\vec{\xi}^{(1)} dV$. Then we must have

$$\frac{\partial}{\partial t} \int_V n f_1 \, dV + \int_S n f_1 \, \vec{\xi}^{(1)} \cdot \vec{\nu} \, dS = \int_V \frac{\partial}{\partial t} n f_1 \Big|_{coll.} dV \ .$$

By applying Gauss theorem to the surface integral and letting $V \to \infty$, it becomes the Boltzmann equation:

$$\frac{\partial}{\partial t} n f_1 + \nabla \cdot n f_1 \vec{\xi}^{(1)} = \frac{\partial}{\partial t} n f_1 \Big|_{coll.} \qquad (IV.1)$$

where $f_1 \equiv f(\vec{\xi}^{(1)})$. In Cartesian coordinates, since $\vec{\xi}^{(1)}$ is a constant vector in the derivation, the Boltzmann equation becomes

$$\frac{\partial}{\partial t} n f_1 + \xi_i^{(1)} \frac{\partial}{\partial x_i} n f_1 = \frac{\partial}{\partial t} n f_1 \Big|_{coll.} \qquad (IV.1)'$$

where x_i are the Cartesian coordinates, $\xi_i^{(1)}$ the Cartesian velocity components of $\vec{\xi}^{(1)}$, and the convention of summing over identical suffixes is adopted. In more abbreviated form, this is sometimes written as

$$D_1 n f_1 = \frac{\partial}{\partial t} n f_1 \Big|_{coll.} \ , \quad D_1 \equiv \frac{\partial}{\partial t} + \xi_i^{(1)} \frac{\partial}{\partial x_i} \ .$$

The evaluation of the collision terms in the right-hand side of the Boltzmann equation of course requires detailed treatment. A molecular model must be chosen first of all. The dynamic processes of collision are usually simplified by making the following assumptions:

(1) only binary collisions occur, and (2) "molecular chaos" prevails. The second assumption means that the joint probability of finding two molecules having $\vec{\xi}^{(1)}$ and $\vec{\xi}^{(2)}$, resp., at a certain place and a certain time is simply the product of the individual probabilities as if the other were absent. Physically, it amounts to the supposition

that except during collisions, the molecules are uncorrelated with each other. Both assumptions are valid so long as $\lambda'/\sigma \gg 1$, the second one in particular having been examined in detail by Jeans (1925, Chapter IV).

We shall only briefly sketch what the collision term looks like. The force field of each molecule is taken to be conservative and spherically symmetric. The binary collision between two molecules of velocity $\vec{\xi}^{(1)}$ and $\vec{\xi}^{(2)}$ turns $\vec{\xi}^{(1)}$ into $\vec{\xi}^{(1)'}$ and $\vec{\xi}^{(2)}$ into $\vec{\xi}^{(2)'}$, and may be represented schematically as

$$(1, 2) \longrightarrow (1', 2') .$$

We refer to this as a "direct collision", causing the loss of molecules "1". Because of the conservative nature of the process, obviously an "inverse collision" can also happen, i.e.,

$$(1', 2') \longrightarrow (1, 2),$$

causing a gain of molecules "1". The total number of either type of collisions must depend upon the available number of the participants, as well as the relative speed Ω between the two molecules and "cross section" S representing the effective target area. Thus, the total number of class "1" molecules lost in dV during time δt, through direct collisions with all possible molecules of class "2" is

$$n f_1 \, dV \, d\vec{\xi}^{(1)} \, \delta t \int_{\vec{\xi}^{(2)}} n f_2 \, \Omega S \, d\vec{\xi}^{(2)} \qquad \text{(IV.2)}$$

f_2 denoting $f(\vec{\xi}^{(2)})$. Similarly the gain of molecules of class "1" in dV during δt, through all inverse collisions involving $1'$ and $2'$ is

$$\int_{\vec{\xi}^{(2)}} n f_1' \, d\vec{\xi}^{(1)'} \, dV \, \delta t \, n f_2' \, \Omega' \, S' \, d\vec{\xi}^{(2)'} .$$

Note that since $(1, 2) \rightarrow (1', 2')$, for given $\vec{\xi}^{(1)}$ the inverse collisions must be summed up over all pairs of the $1'$- and $2'$-molecules through the choice of molecules of different $\vec{\xi}^{(2)}$. However, the details of the collision process show that

$$S' = S, \quad \Omega' = \Omega, \quad d\vec{\xi}^{(2)'} \, d\vec{\xi}^{(1)'} = d\vec{\xi}^{(2)} \, d\vec{\xi}^{(1)} .$$

Thus the gain of class "1" through inverse collision can be recast as

$$\int_{\vec{\xi}^{(2)}} n^2 f_1' f_2' \, \Omega S \, d\vec{\xi}^{(1)} \, d\vec{\xi}^{(2)} \, dV \, \delta t .$$

Consequently, the right-hand side of the Boltzmann equation can be put into a more convenient form and Eq. (IV.1) becomes

$$D_1 n f_1 = \int_{\vec{\xi}(2)} n^2 [f_1' f_2' - f_1 f_2] \Omega S \, d\vec{\xi}^{(2)}. \qquad \text{(IV.3)}$$

In the case of gas in equilibrium, the left-hand side of Eq. (IV.3) vanishes. A possible solution is obtained by setting the integrand in the right-hand side to zero. The procedure amounts to the assumption that each direct collision is exactly balanced by its inverse, often referred to as the "principle of detailed balancing". As applied to Eq. (IV.3), the solution of

$$f_1 f_2 - f_1' f_2' = 0$$

is in general the local Maxwellian distribution $f^{(0)}$, Eq. (III.10), and proved to be unique (Jeans, 1925, pp. 25-28; Grad, 1949). To satisfy the equilibrium requirement, naturally the mean velocity $\vec{U}$ and temperature T here must be independent of space and time.. To justify the "principle of detailed balancing" in this case, it should be mentioned that a consequence of Eq. (IV.3) is "Boltzmann's H-theorem", showing essentially that any distribution should indeed tend to $f^{(0)}$ through the collisions. For further discussions of the theorem, see e.g. Chapman and Cowling (1952, Chapter 4).

It is observed that the expression (IV.2) expresses the total number of collisions involving the class "1" molecules. Therefore we may introduce an average collision frequency Θ_1 for a class "1" molecule, and it must be given by

$$\Theta_1 = \int_{\vec{\xi}(2)} n f_2 \Omega S \, d\vec{\xi}^{(2)}. \qquad \text{(IV.4)}$$

Formally then, the Boltzmann equation may also be written as

$$D_1 n f_1 = -\Theta_1 [n f_1 - n \tilde{f}_1] \qquad \text{(IV.5)}$$

where $\tilde{f}_1$ is to be obtained by identifying the right-hand side with that of Eq. (IV.3), and has the significance of an average distribution function for the outcome of the inverse collisions. The property that $\tilde{f}_1 \to f_1^{(0)}$ as the number of collisions increases should be kept in mind.

We note that according to Eq. (IV.4), the collision frequency is directly

proportional to the number density. In the other limit of $\Theta_1 \to 0$ for extremely rarefied gases the "free molecule flow" is obtained by setting the right-hand side of Eq. (IV.5) to zero, neglecting the collisions completely. The property nf then is propagated without change in the direction $\vec{\xi}^{(1)}$ and at the speed $\left|\vec{\xi}^{(1)}\right|$. By turning Eq. (IV.5) into an integral equation, a first order correction for "near free molecule flows" may be obtained through iteration, using the free molecule solution to evaluate $n\tilde{f}_1$, see, e.g., Willis (1958).

V. The Maxwell Transfer Equations and the Hydrodynamic Equations

The Boltzmann equation is a nonlinear integro-differential equation, evidently very difficult to handle. In flow problems, however, the complete information given by the distribution function is much too detailed and more than necessary. Our interest in most cases is in the average properties of all the molecules within a "fluid element", such as the temperature T and velocity $\vec{U}$. To deduce equations governing these averages we turn to the Maxwell transfer equations. These are obtained by multiplying the Boltzmann equation by any function $Q(\vec{\xi}^{(1)})$ and then integrating over the velocity space $\vec{\xi}^{(1)}$. Since $Q(\vec{\xi}^{(1)})$ is independent of space and time, the result from Eq. (IV.3) is, dropping the superscript (1) on $\vec{\xi}^{(1)}$,

$$\frac{\partial}{\partial t} n\langle Q\rangle + \frac{\partial}{\partial x_i} n\langle Q\xi_i\rangle = \langle \Delta nQ\rangle_{\text{coll.}} \qquad (V.1)$$

The term $\langle \Delta nQ\rangle_{\text{coll.}}$ on the right-hand side is simply an abbreviation of the rather lengthy expression, to be examined immediately.

Written out in full, the term is

$$\langle \Delta nQ\rangle_{\text{coll.}} = \int_{\vec{\xi}^{(1)}}\int_{\vec{\xi}^{(2)}} Q_1 n^2 [f_1' f_2' - f_1 f_2]\, \Omega S\, d\vec{\xi}^{(1)}\, d\vec{\xi}^{(2)},$$

Q_1 standing for $Q(\vec{\xi}^{(1)})$. Since the integration is over all $\vec{\xi}^{(1)}$ and $\vec{\xi}^{(2)}$, the roles of "1" and "2" may be interchanged without affecting the result. Hence, alternatively,

$$\langle \Delta nQ\rangle_{\text{coll.}} = \int_{\vec{\xi}^{(2)}}\int_{\vec{\xi}^{(1)}} Q_2 n^2 [f_2' f_1' - f_2 f_1]\, \Omega S\, d\vec{\xi}^{(2)}\, d\vec{\xi}^{(1)}.$$

We next note that when $\vec{\xi}^{(1)}$ and $\vec{\xi}^{(2)}$ take on all possible values, so do $\vec{\xi}^{(1)'}$ and $\vec{\xi}^{(2)'}$. But,

$$(1,2) \rightleftarrows (1',2').$$

If instead all the inverse processes are considered, the integral may also be expressed in

$$\langle \Delta n Q \rangle_{\text{coll.}} = \int_{\vec{\xi}(1)'} \int_{\vec{\xi}(2)'} Q_1' n^2 [f_1 f_2 - f_1' f_2'] \Omega' S' d\vec{\xi}^{(1)'} d\vec{\xi}^{(2)'}$$

$$= \int_{\vec{\xi}(1)} \int_{\vec{\xi}(2)} Q_1' n^2 [f_1 f_2 - f_1' f_2'] \Omega S \, d\vec{\xi}^{(1)} d\vec{\xi}^{(2)}.$$

Again interchanging the suffixes "1" and "2", we get still another form:

$$\langle \Delta n Q \rangle_{\text{coll.}} = \int_{\vec{\xi}(2)} \int_{\vec{\xi}(1)} Q_2' n^2 [f_2 f_1 - f_2' f_1'] \Omega S \, d\vec{\xi}^{(2)} d\vec{\xi}^{(1)}.$$

Finally, a form symmetrical with respect to the indices "1" and "2" is obtained by using the arithmetic mean of the four equivalent expressions:

$$\langle \Delta n Q \rangle_{\text{coll.}} = \tag{V.2}$$

$$= \frac{1}{4} \int_{\vec{\xi}(1)} \int_{\vec{\xi}(2)} n^2 [f_1' f_2' - f_1 f_2][Q_1 + Q_2 - (Q_1' + Q_2')] \Omega S \, d\vec{\xi}^{(1)} d\vec{\xi}^{(2)}.$$

Eq. (V.2) explicitly shows that if Q is a dynamical property which is a "collisional invariant", i.e.,

$$Q_1 + Q_2 = Q_1' + Q_2'$$

then $\langle \Delta n Q \rangle_{\text{coll.}}$ vanishes. This is, of course, to be expected. The right-hand side of the Boltzmann equation is the net change of the number of molecules with velocity $\vec{\xi}^{(1)}$. When multiplied by Q_1 and summed over all the molecules, the result is the net change of the property Q for the aggregate due to collisions. If the sum of Q does not change in any collision, the total cannot change for all the collisions.

For conservative systems the collisional invariants are mass m, momentum $m\vec{\xi}$ and energy $\frac{1}{2} m\xi^2$. With $Q = m$, Eq. (V.1) gives the "continuity equation" in conventional gasdynamics,

$$\frac{\partial}{\partial t} \rho + \frac{\partial}{\partial x_i} \rho U_i = 0, \tag{V.3}$$

where U_i are the Cartesian components of the mean or fluid velocity $\vec{U}$. With $Q = m\vec{\xi}$, there follows

$$\frac{\partial}{\partial t} \rho \vec{U} + \frac{\partial}{\partial x_i} \rho \langle \xi_i \vec{\xi} \rangle = 0.$$

Let us write

$$\vec{\xi} = \vec{U} + \vec{c}\,, \text{ or in Cartesian components,}$$

$$\xi_i = U_i + c_i\,,$$

where $\vec{c}$ is the "thermal velocity", and obviously $\langle c_i \rangle = 0$. Then the transfer equation for momentum may be recast in Cartesian coordinates after making use of Eq. (V.3), into the conventional form

$$\rho \left[\frac{\partial}{\partial t} U_i + U_i \frac{\partial}{\partial x_j} U_j \right] = \frac{\partial}{\partial x_i} P_{ij} \qquad (V.4)$$

where P_{ij} may be identified with the stress tensor in conventional gasdynamics, and is seen to be given from the molecular viewpoint by

$$P_{ij} = -\rho \langle c_i c_j \rangle\,. \qquad (V.5)$$

If we sum the three "normal stresses" P_{11}, P_{22} and P_{33}, Eq. (V.5) gives

$$P_{ii} = -\rho \langle c_i^2 \rangle = -\rho \langle c^2 \rangle.$$

But $\langle c^2 \rangle = 3/2\beta = 3RT$, and from the perfect gas law $p = \rho RT$. Hence, as is usually defined, we also have

$$p = -\frac{1}{3} P_{ii}\,.$$

After taking the pressure out, the kinetic expression for the "viscous stress tensor" is found to be

$$P'_{ij} = -\rho \langle c_i c_j \rangle + \delta_{ij} p \qquad (V.6)$$

where $\delta_{ij} = 0$ for $i \neq j$, $\delta_{ij} = 1$ for $i = j$.

Similarly, for $Q = m\xi^2/2$, the transfer equation can be manipulated into the form of the conventional "energy equation",

$$\rho c_v \left(\frac{\partial T}{\partial t} + U_i \frac{\partial T}{\partial x_i} \right) = -p \frac{\partial U_i}{\partial x_i} + \Phi + \frac{\partial q_i}{\partial x_i} \qquad (V.7)$$

where Φ is the "dissipation function",

$$\Phi = P'_{ij} \frac{\partial U_i}{\partial x_j}$$

and $\vec{q}$ is the "heat flux vector", kinetically expressed as

$$\vec{q} = -\rho \left\langle \frac{1}{2} c^2 \vec{c} \right\rangle\,. \qquad (V.8)$$

In general Eqs. (V.3), (V.4) and (V.7) are called the "hydrodynamic equations", describing the change of the fluid dynamical properties ρ, $\vec{U}$ and T in terms of the stress tensor and the heat flux vector. If the distribution function

is assumed to be a local Maxwellian, it is easily verified by direct calculation that $P'_{ij} = q_i = 0$, corresponding to the inviscid and non heat-conducting approximation. When deviation from the local Maxwellian is small, we shall see that the Navier-Stokes and Fourier laws emerge. When these are no longer sufficient, there seems to be no other course except through further analysis of the Boltzmann equation to achieve an adequate approximation of the distribution function.

VI. Asymptotic Expansion for Near Maxwellian Distributions

To begin with, let us recall the Boltzmann equation in the form of Eq. (IV.5),

$$D_1 nf_1 = -\Theta_i [\, nf - n\tilde{f}_1 \,] .$$

The left-hand side must have a time constant $\bar{\tau}$ characterizing the change of the function nf_1 as an overall phenomenon. The time constant on the right-hand side is an average of the velocity-dependent collision time $1/\Theta_1$, representing the details. Thus the solution must behave differently depending on the ratio of these two time constants. Let us assume that all Θ_1's are $O(\bar{\Theta})$, an average; the ratio is then $\varepsilon = 1/\bar{\tau}\bar{\Theta}$. When $\varepsilon \ll 1$, it is further expected that $f_1 \cong f_1^{(0)}$.

Restricting to $\varepsilon \ll 1$, a natural procedure therefore is to seek an asymptotic expansion of f_1 in ascending powers of ε:

$$f_1 \sim f_1^{(0)} + \varepsilon f_1^{(1)} + \varepsilon^2 f_1^{(2)} + \cdots . \qquad (VI.1)$$

We now revert to Eq. (IV.3), noting that the left-hand side is $O(\varepsilon)$, compared to the right-hand side and that $f^{(0)}$ is a solution for $\varepsilon = 0$. By substituting Eq. (VI.1) into Eq. (IV.3), to $O(\varepsilon)$ the equation becomes

$$D_1 nf_1^{(0)} = \varepsilon \int_{\vec{\xi}(2)} n^2 [\, (f_1^{(0)'} f_2^{(1)'} + f_1^{(1)'} f_2^{(0)'}) - (f_1^{(0)} f_2^{(1)} + f_1^{(1)} f_2^{(0)}) \,] \Omega S \, d\vec{\xi}(2) .$$

A slightly simpler expression results if we set $\varepsilon f^{(1)} = f^{(0)} \phi$. The equation for ϕ is

$$D_1 nf_1^{(0)} = \int_{\vec{\xi}(2)} n^2 f_1^{(0)} f_2^{(0)} [\, \phi_1' + \phi_2' - \phi_1 - \phi_2 \,] \Omega S \, d\vec{\xi}(2) . \qquad (VI.2)$$

The left-hand side being known, Eq. (VI.2) is a linear integral equation of the Fredholm type. If $f^{(0)}$ is chosen to be indeed the "local Maxwellian",

$$\int f^{(0)}\, d\vec{\xi} = 1, \quad \int f^{(0)} \vec{\xi}\, d\vec{\xi} = \vec{U}, \quad \int f^{(0)} \vec{c}^{\,2}\, d\vec{\xi} = \frac{3}{2\beta} . \qquad (VI.2)$$

Then ϕ must satisfy the following:

$$\int f^{(0)} \phi\, d\vec{\xi} = 0, \quad \int f^{(0)} \phi\, \vec{\xi}\, d\vec{\xi} = 0, \quad \int f^{(0)} \phi\, \vec{c}^{\,2}\, d\vec{\xi} = 0. \qquad (VI.3)$$

Eq. (VI.3) turns out to be sufficient to guarantee a unique solution of Eq. (VI.2) (Chapman and Cowling (1952)). In fact all that is needed eventually is a particular solution satisfying Eq. (VI.3). The result is the famous Chapman-Enskog solution.

The particular solution of course depends on the explicit form of the left-hand side of Eq. (VI.2). Evaluating $D_1 nf_1^{(0)}$ we find,

$$D_1 nf_1^{(0)} = nf_1^{(0)} \left\{ \left(\frac{c^2}{c_m^2} - \frac{5}{2} \right) \vec{c} \cdot \nabla \ln T + b_{ij} \frac{\partial U_i}{\partial x_j} \right\} \qquad (VI.4)$$

where

$$c_m^2 = \frac{1}{\beta}$$

$$b_{ij} = 2 \frac{c_i c_j}{c_m^2} - \frac{2}{3} \frac{c^2}{c_m^2} \delta_{ij} .$$

This expression enables us to be more specific about the time constant $\bar{\tau}$. From the two terms in the bracket, it is seen that

$$\bar{\tau} \sim O\left(\frac{c_m}{\bar{T}} \frac{\Delta T}{L} \right) \quad \text{or} \quad O\left(\frac{\Delta U}{L} \right)$$

where $\bar{T}$ is the characteristic temperature level of the flow, ΔT and ΔU are the temperature and velocity ranges, resp., and L is the characteristic length. In order that $\varepsilon = 1/\bar{\tau}\bar{\Theta} \cong \lambda / c_m \bar{\tau} \ll 1$, we must require $(\lambda/L)(\Delta T/T) \ll 1$ and $(\lambda/L)(\Delta U/c_m) \ll 1$. For a fixed Knudsen number λ/L, the accuracy of the Chapman-Enskog solution improves as $\Delta T/T$ and $\Delta U/c_m$ become smaller.

We shall not go into the details of solving ϕ from Eqs. (VI.2) and (VI.3), for which the reader should consult Chapman and Cowling (1952). It suffices for our purposes to note that, because of the form of Eq. (VI.4), it is possible to represent ϕ by

$$\phi = - A \vec{c} \cdot \nabla \ln T - B b_{ij} \frac{\partial U_i}{\partial x_j} ,$$

A and B being two scalar functions of the thermal velocity $\vec{c}$. Once A and B are determined, by Eqs. (V.6) and (V.8) the viscous stress tensor and the heat flux vector may thus be calculated. Indeed, these assume the same expressions as

the Navier - Stokes and the Fourier laws:

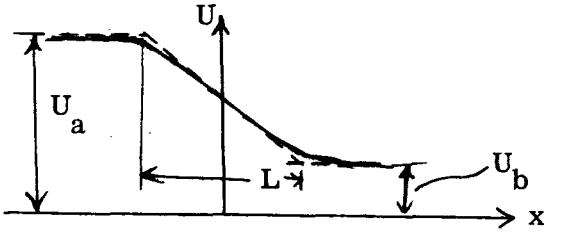
$$P'_{ij} = \mu\left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i}\right) - \frac{2}{3}\mu\frac{\partial U_i}{\partial x_j}$$

$$q_i = k\frac{\partial T}{\partial x_i} \qquad \text{(VI. 5)}$$

and the viscosity coefficient μ and the coefficient of thermal conductivity k are obtained from the functions B and A, resp., through quadrature. The Chapman - Enskog solution, i.e., to $O(\varepsilon)$ only, brings out thereby the restricted validity of the conventional gasdynamics, as well as provides a theoretical means of evaluating μ and k with suitable choices of the molecular model.

When the solution is carried out to $O(\varepsilon^2)$, the details are even more tedious and the resulting formulae for P'_{ij} and q_i are necessarily much more complicated. Because ϕ contains the spatial gradients of temperature and velocity, the left - hand side of the Boltzmann equation now includes second derivatives and products of first derivatives of these mean flow variables. The time derivatives, on the other hand, turn out to be always capable of elimination in favor of the spatial derivatives. The viscous stress tensor and the heat flux vector are thus dependent only on the spatial variations of the mean flow variables. This peculiar set of solutions is sometimes referred to as "normal solutions". When the P'_{ij} and q_i from the solution to $O(\varepsilon^2)$ are substituted into the hydrodynamic equations, the result is known as the "Burnett equations". Since higher order derivatives occur, more boundary conditions than the Navier - Stokes equations are generally required, and therefore to be formulated for the solution of the problem. But, although to $O(\varepsilon^2)$, if Eq. (V. 1) is regarded as an asymptotic expansion, the Burnett equations do not necessarily provide better accuracy.

We shall now apply the Navier - Stokes equations to the one - dimensional steady shock problem, which may be stated as follows: Given a uniform stream of velocity U_a, pressure p_a, density ρ_a, find the solution which permits a smooth transition into another uniform stream downstream. The final velocity U_b, pressure p_b and density ρ_b must



satisfy the conservation laws of mass, momentum and energy, and may thus be regarded as included in the data U_a, p_a, and ρ_a. As is well-known, the relation between U_a and U_b, p_a and p_b, etc. (i.e., between the downstream and upstream quantities) are the Rankine-Hugoniot relations, in which the key parameter is the Mach number,

$$M_a = U_a / \sqrt{\gamma p_a / \rho_a} ;$$

and, to require the entropy change $s_b - s_a > 0$, we must have $M_a > 1$, then $M_b < 1$.

The detailed calculation has been carried out by various authors. Let the attention be focused on a "shock thickness" L defined by the maximum slope and the asymptotic values of one of the flow variables, say U, as sketched. Without explicitly writing out the governing equations, we are content with some general information from simple dimensional reasoning. The gas is characterized by its material properties μ_a and k_a (both are functions primarily of the temperature for a given gas), the flow may be characterized by the prescribed U_a, ρ_a and M_a. But a result of the Chapman-Enskog solution is that μ and k are in fact proportional to each other. Hence the solution must yield

$$L = L(U_a, \rho_a, M_a, \mu_a) .$$

By dimensional homogeneity, it follows

$$L = \frac{\mu_a}{\rho_a U_a} F(M_a)$$

Since $\mu_a \sim \rho_a \lambda_a \bar{c}_a$, the above predicts

$$L \sim \lambda_a F(M_a) .$$

The detailed calculations give $F(M_a) \sim O(1)$ for all finite $M_a > 1$. Thus we conclude that according to the Navier-Stokes equations the shock thickness is $O(\lambda_a)$ -- in other words, most of the changes occur within a distance comparable to the (upstream) mean free path. However, except for weak shocks where the difference between T_a and T_b or U_a and U_b is small, the restriction $\varepsilon << 1$ will be violated, and the significance of the result is open to question.

The Burnett equations have been applied by Zoller (1931) to the shock problem, which happens not to have any ambiguity regarding the boundary conditions. The

selection gives a somewhat larger shock thickness at the lower Mach numbers, but predicts oscillations in the profiles at M_a about 1.3 and breaks down when M_a goes beyond about 2. Available experimental evidences do not support the last two rather peculiar results. (See, e.g., Sherman and Talbot (1960).) As a test case, it is often thought of as an indication that the asymptotic expansion perhaps should not be carried beyond the Chapman - Enskog level.

VII. Methods Based on the Moment Equations

Abandoning the expansion in terms of a small parameter ϵ, we look for alternative ways of finding an approximate solution of the Boltzmann equation, Eq. (IV.3). Similar to the Rayleigh - Ritz or Galerkin methods often used in, e.g., vibration problems, a trial function with a number of adjustable parameters may be assumed, and the latter parameters are to be so chosen that the exact differential equation will be satisfied in some average sense. Now the Maxwell transfer equation, Eq. (V.1), may be interpreted as an average of Eq. (IV.3) in the entire velocity space when the weighing factor is chosen to be Q, including as special cases the hydrodynamic equations, Eqs. (V.3), (V.4) and (V.7). The mean flow variables ρ, U_i, and T may be thus interpreted as the "adjustable parameters" present in the trial function, as indeed also the viscous stress tensor P'_{ij} and the heat flux q_i. The hydrodynamic equations unfortunately are too few in number to determine uniquely all these parameters, except when P'_{ij} and q_i are somehow related to the other parameters, for instance, through the Navier - Stokes and Fourier laws. More equations, of course, could be generated by other choices of Q in Eq. (V.1). On the other hand, for each Q there seems to be no mathematical reason that the Boltzmann equation must be averaged throughout the entire velocity space. If the velocity space is split in two, say $\xi_1 > 0$ and $\xi_1 < 0$, for every one of Eq. (V.1) we get two equations: one from the integration in subspace $\xi_1 > 0$ and one from the integration in subspace $\xi_1 < 0$. All such equations will be referred to as the "moment equations", of which the Maxwell transfer equation itself becomes a special case. By "moment equation method" we mean in general that after assuming a trial function as the approximate solution, the parameters are determined through

the choice of a sufficient number of moment equations of one type or another.

For practical reasons the number of parameters in the trial function will have to be rather limited. The accuracy therefore would be quite profoundly affected by the choice of the trial function and the moment equations. It is obviously desirable to incorporate in the trial function as many as possible of the features of the expected solution. There is yet no guidance on how best to choose the moment equations. But, in contrast to the asymptotic expansion, its applicability is not restricted to any special segment of the entire Knudsen number spectrum.

(A) Grad's Thirteen Moment Equations

As alluded to in the above, by examining the hydrodynamic equations, it would be natural to take the flow variables ρ, U_i, T, as well as P'_{ij} and q_i as the parameters in the trial function. This is precisely what Grad (1949) proposed to do. Because of the symmetry the apparent number of parameters is 14. One of these, however, is redundant, since $p = \rho RT = -(1/3) P_{ii}$. The net number of parameters is therefore 13, and eight more moment equations are needed beyond the hydrodynamic equations. Grad's choice was to use again the Maxwell transfer equations but with $Q = c_i c_j$ and $c_i c^2$. Out of the nine equations that result, one of them is also redundant because the "energy equation", Eq. (V.7), from $Q = c_i c_i$ is already accounted for.

More specifically, the distribution function which Grad took as the trial function is of the form

$$f = f^{(0)} \left[1 + \frac{4}{5} \frac{q_i}{p c_m} \frac{c_i}{c_m} \left(\frac{5}{2} - \frac{c^2}{c_m^2} \right) - \frac{P'_{ij}}{p} \frac{c_i c_j}{c_m^2} \right] \qquad \text{(VII.7)}$$

where $f^{(0)}$ is the local Maxwellian, and $c_m = 1/\sqrt{\beta}$ as before. It is easily verified that the averages $-\rho \langle c_i c_j \rangle$ and $-\rho \langle (1/2) c_i c^2 \rangle$ calculated with Eq. (VII.1) are in agreement with the definitions of P'_{ij} and q_i according to Eqs. (V.6) and (V.8). In fact, it might be remarked that Eq. (VII.1) could be written down directly from the Chapman-Enskog solution by replacing the velocity and temperature gradient terms in the latter with P'_{ij} and q_i through Eq. (VI.5). Generalization of Eq. (VII.1) is possible, as Grad pointed out, by including in the bracket higher order polynomials in c_i orthogonal to the terms present (the

"Hermite polynomials"), amounting to a series expansion of the correction to the local Maxwellian in terms of these polynomials. (Additional parameters and moment equations will then be required.) By limiting to Eq. (VII.1), the deviation from the local Maxwellian therefore is implied to be relatively small.

With $Q = m c_i c_j$, Eq. (V.1) becomes

$$\frac{\partial}{\partial t} \rho \langle c_i c_j \rangle + \frac{\partial}{\partial x_k} \rho \langle c_i c_j \xi_k \rangle = \langle \Delta \rho c_i c_j \rangle_{\text{coll.}}$$

or

$$\frac{\partial}{\partial t} \rho \langle c_i c_j \rangle + \frac{\partial}{\partial x_k} \rho U_k \langle c_i c_j \rangle + \frac{\partial}{\partial x_k} \rho \langle c_i c_j c_k \rangle = \langle \Delta \rho c_i c_j \rangle_{\text{coll.}}$$

The left-hand side averages can also be evaluated with Eq. (VII.1) but the right-hand side depends on the molecular model. The most convenient model is the so-called Maxwell molecule, which repels another like molecule with a force proportional to r^{-5}, r being the distance between the two molecules. For such molecules, the equation finally may be written as

$$\begin{aligned} &\frac{\partial}{\partial t} P'_{ij} + \frac{\partial}{\partial x_k} U_k P'_{ij} + \frac{2}{5} \left(\frac{\partial q_i}{\partial x_j} + \frac{\partial q_j}{\partial x_i} - \frac{2}{3} \delta_{ij} \frac{\partial q_k}{\partial x_k} \right) \\ &\quad + P'_{ik} \frac{\partial U_j}{\partial x_k} + P'_{jk} \frac{\partial U_i}{\partial x_k} - \frac{2}{3} \delta_{ij} P'_{k\ell} \frac{\partial U_k}{\partial x_\ell} \\ &\quad - p \left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} - \frac{2}{3} \delta_{ij} \frac{\partial U_k}{\partial x_k} \right) = -\frac{p}{\mu} P'_{ij} \end{aligned} \qquad \text{(VII.2)}$$

where μ denotes the expression that gives the viscosity coefficient according to the Chapman-Enskog solution for the same molecular model. Likewise, with $Q = (1/2)\, m c_i c^2$, Eq. (V.1) leads to

$$\begin{aligned} &\frac{\partial q_i}{\partial t} + \frac{\partial}{\partial x_j} U_j q_i + \frac{7}{5} q_j \frac{\partial U_i}{\partial x_j} + \frac{2}{5} \left(q_j \frac{\partial U_j}{\partial x_i} + q_i \frac{\partial U_j}{\partial x_j} \right) \\ &\quad + RT \frac{\partial P'_{ij}}{\partial x_j} + \frac{7}{2} P'_{ij} \frac{\partial RT}{\partial x_j} - \frac{P'_{ij}}{\rho} \left[\frac{\partial P'_{jk}}{\partial x_k} - \delta_{jk} \frac{\partial p}{\partial x_k} \right] \\ &\quad - \frac{5}{2} p \frac{\partial RT}{\partial x_i} = -\frac{2}{3} \frac{p}{\mu} q_i \,. \end{aligned} \qquad \text{(VII.3)}$$

From Eqs. (VII.2) and (VII.3), we see that in Grad's solution, there is considerable interaction between the stress tensor and the heat flux. More striking when compared with the Chapman-Enskog solution is the presence of the explicit time derivative term in both equations. Thus, if there are no spatial variations, the equations reduce to

$$\left.\begin{aligned} \frac{\partial}{\partial t} P'_{ij} &= -\frac{p}{\mu} P'_{ij} \\ \frac{\partial}{\partial t} q_i &= -\frac{2}{3}\frac{p}{\mu} q_i \end{aligned}\right\} .$$

A relaxation phenomenon, non-existent in the Chapman-Enskog solution, is now predicted. The time constant is $O(\mu/p)$. Since $\mu \sim \rho \bar{c} \lambda$,

$$\frac{\mu}{p} \sim \frac{\rho \bar{c} \lambda}{\rho RT} \sim \frac{\lambda}{\bar{c}} \sim \frac{1}{\Theta}$$

which is of course the expected order of magnitude following Eq. (IV.5) as discussed in the previous section. But here the result is more quantitative and in particular P'_{ij} and q_i are found to have somewhat different relaxation times.

Of considerable interest is the fact that both the Chapman-Enskog and Burnett formulae for P'_{ij} and q_i can be obtained from the Grad equations, even though the Grad equations are obtained from a distribution function, Eq. (VII.1), that is at the level of the Chapman-Enskog solution only. This is done by regarding Eqs. (VII.2) and (VII.3) as definitions for the right-hand side quantities, namely, P'_{ij} and q_i. If P'_{ij} and q_i are small, the effects of the presence of these quantities in the left-hand side may be determined through an iteration process, starting from $P'_{ij} = q_i = 0$. The first iteration then gives precisely the Navier-Stokes and Fourier laws. The second iteration gives the Burnett result, after eliminating $(\partial U_i/\partial t)$ and $(\partial T/\partial t)$ by means of the hydrodynamic equations. This feature of the Grad equations is the effort in achieving the Chapman-Enskog, not to say the Burnett, solution, it also demonstrates the power of the moment equation method when properly used.

We skip over the question of the boundary conditions for the Grad equations, which have been examined to some extent by Grad himself. As applied to the one-dimensional steady shock problem (Grad (1952)) at lower Mach numbers these equations yield solutions which are rather close to the Navier-Stokes result, giving a slightly larger shock thickness; but for Mach numbers greater than about 1.65, again no solution can be found. This is to some extent rather disappointing. The difficulty could only be attributed to the chosen form of Eq. (VII.1), which ceases to provide a good approximation when the molecules are far from being in a state

of quasi-equilibrium. For lower Mach numbers, i.e., weak shocks, the up- and downstream conditions are not too different from each other, the distribution anywhere within the shock thus deviates little from an average constant Maxwellian. Such is, of course, far from being the case for large Mach numbers and stong shocks.

(B) Mott-Smith's Bimodal Distribution

We now recognize that for strong shocks a trial function not restricted to quasi-equilibrium is necessary. A very simple choice was offered by Mott-Smith (1951), who assumed that it might be taken as a linear combination of the up- and downstream distribution functions,

$$f = \alpha_a(x) f_a^{(0)} + \alpha_b(x) f_b^{(0)} \qquad \text{(VII.4)}$$

where $\alpha_a(x)$ and $\alpha_b(x)$ are the adjustable parameters. However, since $\int f\, d\vec{\xi} = \int f_a^{(0)} d\vec{\xi} = \int f_b^{(0)} d\vec{\xi} = 1$, we require

$$\alpha_a + \alpha_b = 1, \qquad \text{(VII.5)}$$

If the x-axis is in the direction of flow, the boundary conditions are

$$\begin{aligned} &x \to -\infty, \quad \alpha_a \to 1, \quad \alpha_b \to 0; \\ &x \to +\infty, \quad \alpha_a \to 0, \quad \alpha_b \to 1. \end{aligned} \qquad \text{(VII.6)}$$

The "bimodal" nature is clear, as for given x the molecules may be regarded as a mixture of two groups maintaining either the up- or downstream characteristics.

There is now in effect only one adjustable parameter. To determine this parameter, Mott-Smith left the hydrodynamic equations alone but employed a moment equation obtained from Eq. (V.11) with $Q = \xi_1^2$, or ξ_1^3, which then was solved by imposing Eq. (VII.6). The hydrodynamic equations provide as usual the Rankine-Hugoniot relations, expressing all downstream properties in terms of those upstream. A solution for α_a, say, at all $M_a > 1$ was shown to be possible and the flow variables computed as averages. The shock thicknesses so determined from the two choices of Q differ between themselves by 10 to 25 percent depending on M_a. This difference, of course, reflects the uncertainty due to the arbitrariness in choosing Q. There have been consequently discussions attempting to arrive at a criterion for the selection (e.g., Rosen (1954), Sakurai (1951)). More realistic molecular models have also been used in evaluating the collision terms of the moment

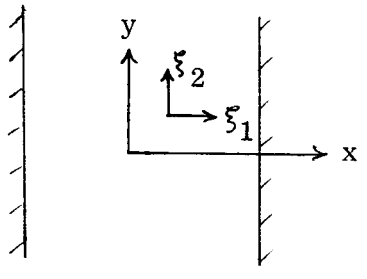
equation (Muckenfuss (1960)). At the lower Mach numbers, the Mott-Smith shock thickness is much greater than that from the Navier-Stokes or Grad equations, and generally considered inaccurate. A comparison is shown in the accompanying figure.

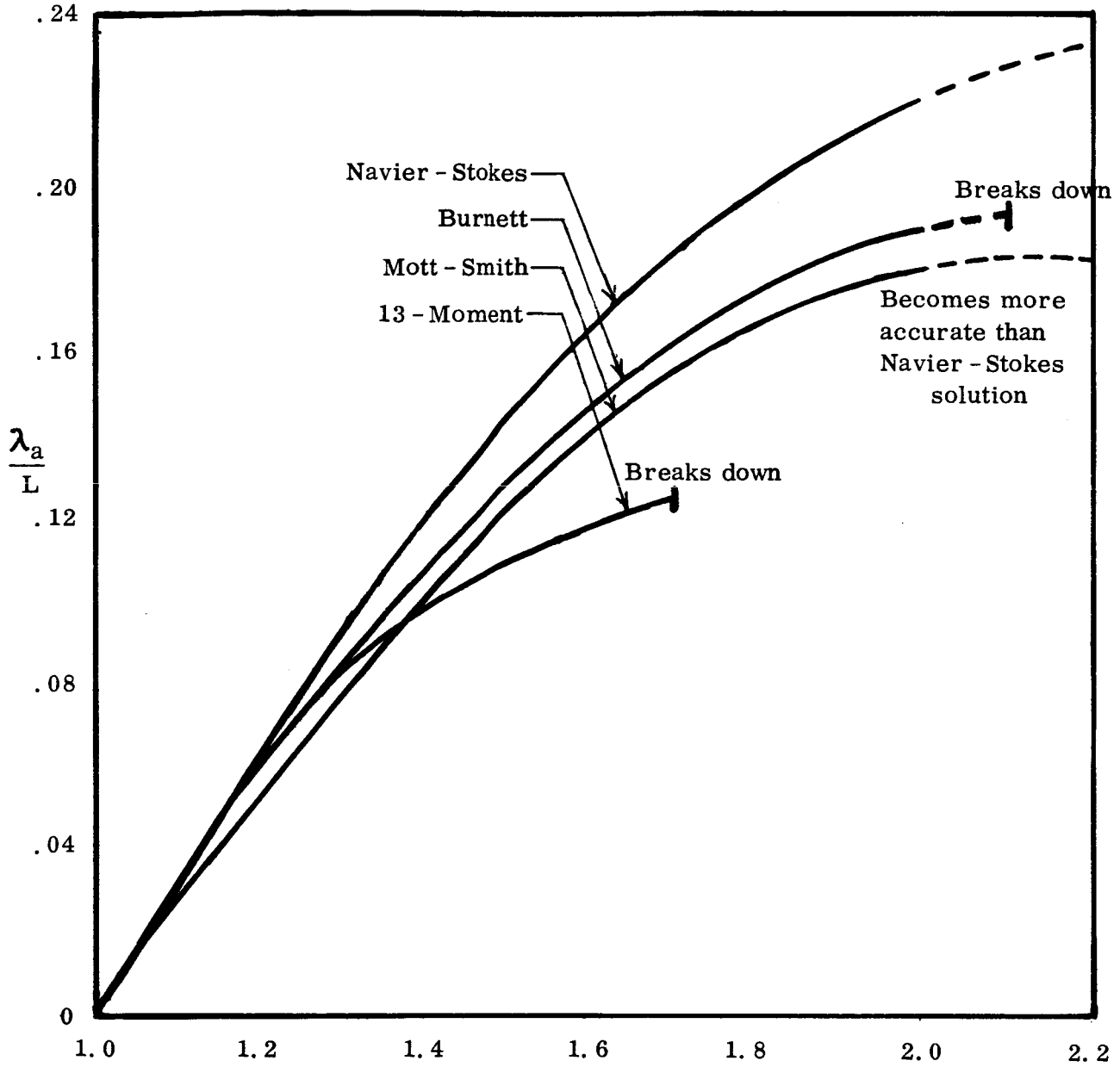
A point worth noting, however, is that with a single adjustable parameter the Mott-Smith distribution function can hardly be expected to be accurate in the details. Since ρ, U, etc. are found by averaging Eq. (VII.4), the hydrodynamic equations are not satisfied anywhere within the shock, except for the finite changes between the up- and downstream conditions. An alternative avoiding this difficulty seems to be that the approximate distribution function might be used only for the viscous stress and heat flux terms needed in the hydrodynamic equations, which then can be solved in much the same manner as with the Navier-Stokes equations. The result will also necessarily satisfy the Rankine-Hugoniot relations.

(C) Methods Using Half-Range Distribution Function

We have mentioned that the shock wave is a convenient example in rarefied gasdynamics because of the absence of solid boundaries. When a solid boundary is present, the molecules rebound from, or rather are emitted by the solid boundary, and usually have "forgotten" most of their past history. In fluid elements near to the solid boundary, therefore, the distribution function would be discontinuous in the velocity space, in the sense that those moving toward the solid boundary and those going away from it would require quite different expressions. An expansion of the Grad type in terms of continuous functions, Eq. (VII.1) for instance, would need a very large number of terms to approximate a discontinuity adequately. If the discontinuous nature is recognized beforehand and taken care of separately, however, the remainder would be much easier to approximate.

This observation was exploited by Gross and Ziering (1958) in their investigation of several problems involving the geometry of two parallel plates when the gas in between may be highly rarefied. Let the direction normal to the two plates be x. The molecules are assumed to be





COMPARISON OF RECIPROCAL SHOCK THICKNESS
FOR MONATOMIC MAXWELL MOLECULES

(From Ziering, S., Ek, F., & Koch, P., Phy. Fluids, 4, 975-987, 1961)

composed of two groups according to the sign of ξ_1. Then for the distribution function we write

$$nf = n_{\pm} f_{\pm} \quad \text{for} \quad \xi_1 \gtrless 0 \qquad \text{(VII.7)}$$

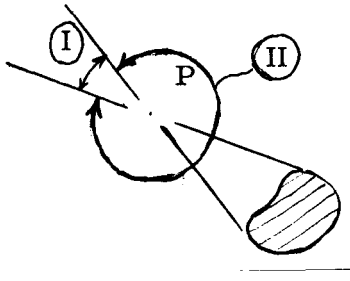
where $n_{\pm}$ are functions of x, and $f_{\pm}$ are defined only in the half spaces $\xi_1 \gtrless 0$, resp., hence referred to as half-range distribution functions. The functions $f_{\pm}$ used by Gross and Ziering are expressed in terms of the Hermite polynomial, similar to Eq. (VII.1), slightly modified because the orthogonality condition now is to be applied in the half spaces. Compared with Grad's approach, with the same number of parameters in the expansion, evidently the half-range distribution function contains twice as many unknowns; consequently, twice as many moment equations are needed for their determination. Gross and Ziering then split up each of the Maxwell transfer equations, Eq. (V.1), into two by carrying out the integration in the two halves of the velocity space separately. In a Grad-like expansion of the half-range distribution function, the adjustable parameters lose the physical significance as corresponding to P'_{ij}, q_i, etc., which are, by definition, the averages over the whole velocity space.

Application of the technique has been limited to sevaral "linearized" problems where the relative velocity or the temperature difference of the two plates is small. In such cases, the half-range distributions were expanded around a constant Maxwellian, and the calculation was rather straightforward.

An alternative choice of the half-range distributions in Eq. (VII.7) is the "two-stream Maxwellian" distribution proposed by Lees (1959). The form is taken to be

$$f_{\pm} = \left(\frac{\beta_{\pm}}{\pi}\right)^{3/2} \exp\left[-\beta_{\pm}\left(\vec{\xi} - \vec{u}_{\pm}\right)^2\right] \qquad \text{(VII.8)}$$

where $\beta_{\pm}$ and $\vec{u}_{\pm}$ are the adjustable parameters, in addition to $n_{\pm}$, to be determined by the moment equations. In fact, to generalize the method Lees adopts a "line of sight principle" which divides the molecules into groups as if in free molecule flow. In the problem of the gas between parallel plates, there are thus the same two groups in the half-range representation of Gross and Ziering, each moving toward one of the walls. For the case of an arbitrary body moving in an unbound gas region, at the given point P a pencil of rays may be drawn to form a cone tangent to the body. These molecules in a volume element at P



coming from the body, if in free molecule flow, would have their velocity lying within the cone (I). These are taken by Lees as group I. The rest are all taken as group II. The "dual-range" character of the distribution is then expressible in the same form as Eqs. (VII.7) and (VII.8), except replacing suffices "$\pm$" by suffices "I" and "II". For the needed moment equations Lees prefers to maintain the whole-range transfer equations, Eq. (V.1), with successive Q's similar to Grad's that led to the thirteen moment equations. In contrast to Grad's distribution, there are now, however, only ten parameters $n_{I,II}$, $\beta_{I,II}$ and $\vec{u}_{I,II}$, none of which, it may be noted, has any significance as physical observables except in dimension. Regarding the transfer equations corresponding to Eqs. (VII.2) and (VII.3) as expressions for P'_{ij} and q_i in terms of the left-hand side terms, we are inclined to conclude that, together with the hydrodynamic equations, there still should be thirteen equations for the thirteen variables $(\rho, U_i, T, P'_{ij}, q_i)$ at this level of approximation. In general the ten parameters inherent in the "two-stream Maxwellian" appear too few in number to be really self-consistent. When applied to the parallel plates problem with large relative velocity, some difficulty was indeed experienced by Lees and Lin (1961). Their mention of the possible improvement by using skewed "two-stream Maxwellians" amounts to an effort toward additional degrees of freedom.

A further point of criticism may be directed at the "line of sight" principle. The grouping of molecules following this principle is of course correct in the free molecule limit or very close to the body surface, but the principle seems to be rather irrelevant after the molecules have gone through several collisions. Its consequences therefore need not agree with the result from the Navier-Stokes equation in the conventional continuum limit. This drawback is illustrated in the problem of the cylindrical Couette flow (in the annulus between two rotating concentric circular cylinders) investigated by Ai (1960).

In spite of these objections, the "two-stream Maxwellian" is relatively easy to work with and together with the "line of sight" principle can be used to set up, at least formally, the governing equations for flows involving arbitrary geometry and large deviations from quasi-equilibrium. It would be of interest to see the solution of the shock problem by this method, which unfortunately is not available.

VIII. The BGK Model Equation and the Shock Solution

We have discussed above some of the approximate methods of handling the Boltzmann equation. An entirely different approach is to try for exact solutions by simplifying the Boltzmann equation itself. The most well known of such simplifications is the BGK or Krook model (Bhatnagar et al (1954), Krook (1955)). Looking back, we have the Boltzmann equation, Eq. (IV.5),

$$\mathcal{D}_1 \, nf_1 = -\Theta_1 \, [\, nf_1 - n\tilde{f}_1 \,] .$$

The complications are all contained in the right-hand side terms, which will now be approximated.

First of all, the dependence of Θ_1 on the molecular velocity $\vec{\xi}^{(1)}$ is clearly a matter of detail. It seems reasonable to approximate it with simply $\bar{\Theta}(\vec{r}, t)$, an average for all molecules. To simplify the very complicated $\tilde{f}_1$, the choice is made so as to preserve the following important properties of the exact equation:

(a) As $\bar{\Theta} \to \infty$, $f \to f^{(0)}$, the local Maxwellian.

(b) In the transfer equation, Eq. (V.1), $\langle \Delta n Q \rangle_{coll.} = 0$ for the collisional invariants $Q = m$, $m\vec{\xi}$, $mc^2/2$.

(c) There is an "H-theorem".

Krook took directly $\tilde{f} = f^{(0)}$; the model equation is thus

$$\mathcal{D}_1 nf_1 = -n\Theta_1 (f_1 - f_1^{(0)}) . \qquad \text{(VIII.1)}$$

That the right-hand satisfies the requirements (a) and (b) is immediately obvious. It can be shown that condition (c) is also fulfilled. The rate of change of the function nf is now proportional to its departure from the quasi-equilibrium distribution $f^{(0)}$. Hence Eq. (VIII.1) may be regarded as a relaxation model. The equation is, however, only apparently linear, since the parameters ρ and $\vec{U}$ in $f^{(0)}$ remain to be averaged over the unknown f.

All the previous approximate methods of treating the Boltzmann equation can, of course, be applied to Eq. (VIII.1). The Chapman-Enskog type of solution, for instance, is obtained by writing

$$f = f^{(0)} + \varepsilon f^{(1)} + \cdots .$$

By substitution into Eq. (VIII.1), the solution for $\varepsilon f^{(1)}$ is explicitly given as, dropping subscript "1",

$$\varepsilon f^{(1)} = \frac{1}{n\Theta} \mathcal{J} nf^{(0)} \quad . \tag{VIII.2}$$

The dependences on the mean flow gradients $\nabla \ln T$ and $\partial U_i / \partial x_j$ follow from the same term $\mathcal{J} nf^{(0)}$ as in the Chapman - Enskog solution. The Navier - Stokes and Fourier laws are consequently recovered, except that the viscosity coefficient and the coefficient of thermal conductivity are more crudely predicted.

The thirteen moment equations of Grad can also be derived for Eq. (VIII.1). The left - hand sides of Eqs. (VII.2), (VII.3) are unchanged if Eq. (VII.1) is maintained. The right - hand sides depend on the details of collisions but with Eq. (VIII.1) they can be written down by inspection. The counterparts to Eqs. (VII.2) and (VII.3) are thus found to be

$$\left.\begin{aligned} \frac{\partial}{\partial t} P'_{ij} + A_{ij} &= -\bar{\Theta} P'_{ij} \\ \frac{\partial}{\partial t} q_i + B_i &= -\bar{\Theta} q_i \end{aligned}\right\} \tag{VIII.3}$$

where A_{ij} and B_i stand for all the terms except the time derivative in the left - hand sides of the corresponding Grad equations. The only difference lies in replacing the two relaxation times μ/p and $(3/2)(\mu/p)$ with a single time constant $1/\bar{\Theta}$. For this reason, the Krook approximation is sometimes referred to as the single relaxation model. The comparison also suggests that the average $\bar{\Theta}$ may be taken to be

$$\bar{\Theta} = \frac{p}{\mu} \quad \text{or} \quad \frac{2}{3}\frac{p}{\mu} \tag{VIII.4}$$

depending on whether P'_{ij} or q_i is the dominant feature.

In the near continuum regime which is adequately described by the Grad equations, the difference between the BGK model and the Boltzmann equation amounts thus to a difference in the Prandtl number $Pr \equiv \mu C_p / k$. The correct value is $2/3$ for monatomic gases while from the BGK model the Prandtl number will be unity. The regime of free molecule flow in the limit $\Theta_1 \to 0$ is unaffected by the approximation. Its validity in the transition regimeis rather difficult to assess, although the common belief is that it should serve as a reasonable interpolation.

The integral equation form of Eq. (VIII.1) has been the basis for a number of applications. For brevity consider the steady flow problem:

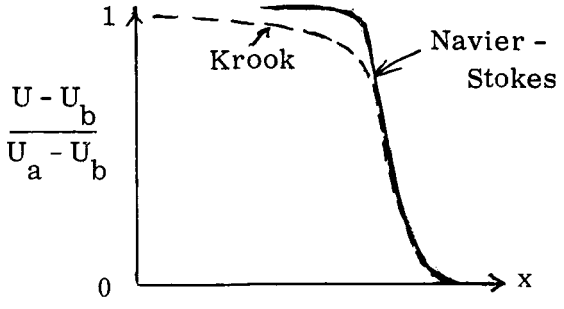
$$\xi^{(1)} \frac{d}{ds_1} nf_1 = -\bar{\Theta}\, n \left[f_1 - f_1^{(0)} \right] \qquad \text{(VIII. 5)}$$

where s_1 is the distance along the direction of $\vec{\xi}^{(1)}$. Direct integration yields, after dropping the subscript "1",

$$nf = n'f' \exp\left[-\int_{s'}^{s} \bar{\Theta}\, ds/\xi \right] + \exp\left[-\int_{s'}^{s} \bar{\Theta}\, ds/\xi \right] \int_{s'}^{s} nf^{(0)} \exp\left[\int_{s'}^{s} \bar{\Theta}\, ds/\xi \right] \bar{\Theta}\, ds/\xi \qquad \text{(VIII. 6)}$$

where the boundary condition $n'f'$ at $s = s'$ is assumed given. The integral $\int_{s'}^{s} \bar{\Theta}\, ds/\xi$ represents the number of collisions for such molecules in traveling the distance between s and s', and the exponential factor is the probability that the molecules from s' should survive. The second term is the gain of such molecules as collision products. Since the unknown functions n and f are involved in the latter, usually an iterative procedure is necessary to achieve a solution. For near free molecule or near continuum flows, a good initial approximation of $nf^{(0)}$ is immediately available. For the transition regime, the initial approximation may have to be found by first doing a cruder analysis, such as the Lees method discussed before. Several problems of the flow between two parallel plates are thus solved by this method (e.g., Willis (1962)).

In applying to the shock structure problem, since the Navier-Stokes solution is reliable for the lower Mach numbers, and obtainable for any shock strength, it becomes an obvious choice as the initial approximation. Such was suggested by Burgers (1956) in his analysis of the problem, but without actually carrying out the calculation. Recently Liepmann et al. (1962), apparently independently, solved the problem by the same procedure in a computer, taking $\bar{\Theta} = p/\mu$, i.e., $Pr = 1$. The solution shows no anomaly at least for Mach numbers as high as 10. A typical comparison against the Navier-Stokes solution is schematically as shown. The agreement with the Navier-Stokes profile is very close in the downstream half, but the upstream portion is considerably more spread out, especially at the higher Mach numbers. This is understandable since the effective coordinate is really $\int \bar{\Theta}\, dx$, so



the physical distance should be inversely proportional to the collision frequency, hence the density, which value for the upstream portion is a small fraction of that for the downstream portion for the stronger shocks.

The Liepmann solution of the shock structure based on the BGK model is unquestionably the most satisfactory to date. It is, however, only the exact solution of an approximation to the Boltzmann equation.

IX. Further Discussion of the Approximate Solution of the Boltzmann Equation

Attractive as the BGK model is, we must not lose sight of the fact that it does not replace the Boltzmann equation. Better and better approximate methods presumably could be developed for the exact equation, and some of them would eventually surpass the BGK model in accuracy. Thus, we return to a discussion of the possible improvement in approximate methods, especially from the viewpoint that these should be applicable throughout the range from continuum to free molecule flow, as the BGK model is.

To gain some perspective, consider the simple differential equation

$$\left.\begin{aligned} \varepsilon \frac{df}{dt} &= -f + f^{(0)}(t) \\ f(0) &\ \text{given} \end{aligned}\right\} \qquad \text{(IX.1)}$$

where ε is a constant parameter of arbitrary magnitude.* Eq. (IX.1) evidently embodies the most important features of the Boltzmann equation, ε having the same significance as $1/\bar{\Theta}$. It is also an analog of the BGK model. The boundary condition $f(0)$ corresponds to the known distribution at an initial instant, or the boundary $s = s'$ as in Eq. (VIII.6). The solution of Eq. (IX.1) can be immediately written down:

$$f = f(0)\, e^{-t/\varepsilon} + \frac{1}{\varepsilon} e^{-t/\varepsilon} \int_0^t f^{(0)} e^{t/\varepsilon}\, dt \qquad \text{(IX.2)}$$

which may be regarded as the simplified version of Eq. (VIII.6).

If t is kept fixed and finite, the asymptotic solutions for $\varepsilon \to \infty$ or $\varepsilon \to 0$ are easily obtained from Eq. (IX.2). For $\varepsilon \to \infty$, the result is

* I am indebted to Prof. G. S. S. Ludford in calling attention to a very similar example in Erdelyi's paper on singular perturbations (Atti Accad. Sci. Torino 95 (1960-61), 651-672).

$$f \cong f(0) + \frac{1}{\varepsilon}\int_0^t [f^{(0)} - f(0)]\, dt + O\left(\frac{1}{\varepsilon^2}\right), \qquad (IX.3)$$

while for $\varepsilon \to 0$, we get

$$f \cong f(0) - \varepsilon \frac{df^{(0)}}{dt} + O(\varepsilon^2). \qquad (IX.4)$$

The first term of Eq. (IX.3) corresponds to the "free-molecule" flow approximation, and the second term is the equivalent of the "first collision" correction usually obtained by one iteration from the free-molecule solution. In the same analogy, the first term of Eq. (IX.4) corresponds to the local Maxwellian in quasi-equilibrium, and the second term is the counterpart of the Chapman-Enskog correction. Both types of asymptotic solutions, as we now see, are not uniformly valid for all t. (Note, in particular, that Eq. (IX.4) can never satisfy the prescribed boundary condition $f(0)$.) For a given ε however large, there is an upper limit of t beyond which the "free-molecule" type of asymptotic expansion ceases to be valid. In the other limit, for a given ε however small, there is a lower limit of t below which the asymptotic expansion for small ε is of no value. Although elementary, this demonstration seems to focus on some of the basic properties of the Boltzmann equation. The pitfalls of trying to push either type of asymptotic expansions into the "transition regime", where $\varepsilon \sim O(1)$, are thus obvious.

The situations for $\varepsilon \to \infty$ and $\varepsilon \to 0$ are very similar to the problems of finding asymptotic solutions for low and high Reynolds numbers, resp., in viscous flow theory. Corresponding to the limit of $\varepsilon \to \infty$, the "Stokes theory" for very low Reynolds numbers is known to be involved in an unbound fluid at sufficient distances from the body. In the other limit of $\varepsilon \to 0$, the analogy to the Knudsen layer is the boundary layer near the body in conventional gasdynamics. The boundary layer thickness goes down as the viscosity is decreased. For points at fixed distances from the body surface, they will eventually lie in the effectively inviscid portion of the flow if the viscosity is small enough. To see the details in the boundary layer, the point in question must be made to move closer to the body surface as the viscosity is reduced, in order to remain within the boundary layer; then and only then the limit for vanishing viscosity may be taken. If we change the words "boundary layer" to "Knudsen layer" and "viscosity" to "mean free path", the last three sentences describe exactly what should be done for analysis of the Knudsen layer as $\varepsilon \to 0$.

As also disclosed by the exact solution, Eq. (IX.2), the natural independent variable should indeed be $\tilde{t} = t/\varepsilon$. We now keep $\tilde{t}$ fixed and finite but let $\varepsilon \to 0$. If Eq. (IX.2) is expressed in $\tilde{t}$ and then integrated by parts, the resulting expansion is found to be

$$f \cong f(0)\, e^{-\tilde{t}} + [\, f^{(0)} - \varepsilon \frac{df^{(0)}}{dt}\,] - e^{-\tilde{t}} [\, f^{(0)}(0) - \varepsilon \frac{df^{(0)}}{dt}\Big|_{t=0}\,] + O(\varepsilon^2).$$

By expanding $f^{(0)}$ and $df^{(0)}/dt$ for small ε two alternative forms, both accurate to $O(\varepsilon)$ <u>for finite $\tilde{t}$</u>, are obtained,

$$f \cong f(0)\, e^{-\tilde{t}} + [\, f^{(0)}(0) - \varepsilon \frac{df^{(0)}}{dt}\Big|_{t=0}\,]\,[\,1 - e^{-\tilde{t}}\,] + \varepsilon \frac{df^{(0)}}{dt}\Big|_{t=0} \tilde{t} + \cdots \tag{IX.5}$$

$$f \cong f(0)\, e^{-\tilde{t}} + [\, f^{(0)} - \varepsilon \frac{df^{(0)}}{dt}\,]\,[\,1 - e^{-\tilde{t}}\,] - \varepsilon \frac{df^{(0)}}{dt}\Big|_{t=0} \tilde{t}\, e^{-\tilde{t}} + \cdots. \tag{IX.5'}$$

The second form is clearly preferable since it remains valid as $\tilde{t} \to \infty$, merging smoothly into the "outer expansion" Eq. (IX.4) and taking on the prescribed boundary value at $t = 0$.

It is now possible to estimate more accurately the "thickness" t_K of the Knudsen layer, in the sense that beyond which, to $O(\varepsilon)$, Eqs. (IX.5') and (IX.4) agree with each other. The condition is therefore

$$e^{-\tilde{t}_K} \sim O(\varepsilon^2)$$

i.e.,

$$\tilde{t}_K \sim O(\ln \varepsilon)$$

or $t_K \sim O(\varepsilon \ln \varepsilon)$. In fact, no matter to what finite order of ε is expanded the asymptotic solution Eq. (IX.4), the same argument will show that t_K is always $O(\varepsilon \ln \varepsilon)$. The Knudsen layer remains to be treated separately.

Returning to the central problem of formulating an approximate solution for the Boltzmann equation, we suggest that the distribution function should exhibit much the same basic features as the solution Eq. (IX.2) of the simplified model. In a moment equation approach, for instance, a reasonable choice of the trial function might resemble Eq. (IX.5'). If we assume an average collision frequency $\bar{\Theta}$ as in the BGK model, a convenient form is

$$nf = n'f' \exp\Big[-\int_{s'}^{s} \bar{\Theta}\, ds/\xi\Big] + n_0 f_0 \Big[1 - \exp\Big[-\int_{s'}^{s} \bar{\Theta}\, ds/\xi\Big]\Big] \tag{IX.6}$$

where $n_0 f_0$ is chosen to contain the adjustable parameters to be controlled by satisfying the moment equations. Since we work with only the average properties

ρ, U_i, P'_{ij}, q_i in the moment equations, there is considerable leeway in the choice of $n_0 f_0$, the main restriction being that it must reproduce the Navier - Stokes and Fourier laws in the limit of $\bar{\Theta} \to \infty$. The free - molecule behavior is guaranteed by the $n'f'$ term, which automatically divides the molecules into groups depending on their "origin". No further assumption such as the Lees "line of sight" principle is now necessary.

It may be noted that the term analogous to the last one in Eq. (IX. 5') is omitted in Eq. (IX. 6) for brevity. The effect presumably is comparable to the net difference from alternative choices of $n_0 f_0$. It is to be emphasized, however, that Eq. (IX. 6) is not meant to be so assessed. Only the form is suggested by the BGK model. The approximation itself is adjusted to satisfy the moment functions of the exact Boltzmann equation, and any molecular model may be adopted for evaluating the collision integrals.

A source of difficulty in the use of Eq. (IX. 6) is the concept of an average collision frequency $\bar{\Theta}$. As in the discussion of the BGK model following Eq. (VIII. 3), the choice of $\bar{\Theta}$ appears to be either p/μ or $(2/3)(p/\mu)$. Besides this ambiguity, any choice of a single average $\bar{\Theta}$ of course over - estimates the mean free path of the fast - moving molecules, as shown by the smaller numerical factor 2/3 needed for matching the heat transfer by means of the BGK model. In assuming Eq. (IX. 6), on the other hand, we have effectively used the BGK model to suggest a way of grouping the free - molecule - like and Navier - Stokes - like molecules. Thus there is no strong reason not to allow $\bar{\Theta}$ to vary somewhat with the speed of ξ, thereby compensating for this source of error. The refinement, however, may or may not be worthwhile, because, again, the net difference might be comparable to that from the alternative choices of $n_0 f_0$. In other words, the nature of the approximation Eq. (IX. 6) is, as a first step and like the BGK model, only to guarantee a smooth transition between the free - molecule and the Navier - Stokes limits. The resulting macroscopic equations are already rather cumbersome to attack, and have been solved only for the simple cases of the linearized plane and cylindrical Couette flows (Shen (1963)). It seems yet premature to introduce further complications.

To conclude this brief survey of the current status of rarefied gasdynamics, we reiterate that our emphasis has been on the treatment of flow problems in terms of

the observables such as mean velocity, pressure, temperature, shear stress and heat flux. The aim is thus essentially to look for the replacement of the conventional Navier - Stokes and Fourier relations in the hydrodynamic equations of motion, applicable throughout the entire range of Knudsen numbers. It might be said that to various degrees of approximation methods are indeed slowly emerging. Unfortunately the geometry of the problem will always enter into resulting equations, so in effect special attention is required for each class of problems defined by its geometry. These equations furthermore are much more complicated than the Navier - Stokes, and our experiences are still confined to the simplest possible examples. The shock wave structure, because of its independence from solid boundaries, has been one of the ideal testing grounds for workers in this rapidly advancing field.

REFERENCES

Ai, D. K. (1960), Calif. Inst. Tech. GALCIT Hypersonic Research Proj. Memo. 56.

Bhatnagar, P. L., Gross, E. P., and Krook, M. (1954), Phy. Rev. 94, 511 - 525.

Burgers, J. M. (1956), Univ. of Md., Inst. Fluid Dyn. Appl. Math.

Chapman, S. and Cowling, T. G. (1952), "The Mathematical Theory of Non - Uniform Gases", 2nd edition, Cambridge University Press.

Grad, H. (1949), Comm. Pure Appl. Math. 2, 331 - 407.

Grad, H. (1952), Comm. Pure Appl. Math. 5, 257 - 300.

Gross, E. P. and Ziering, S. (1958), Phy. Fluids 1, 215 - 224.

Jeans, J. (1925), "The Dynamic Theory of Gases", 4th edition, Cambridge Univ. Press, also Dover Publications.

Krook, M. (1955), Phy. Rev. 99, 1896 - 1897.

Lees, L. (1959), Cal. Inst. Tech. GALCIT Hypersonic Res. Proj. Memo. 51.

Lees, L. and Lin, C. Y. (1961), "Rarefied Gas Dynamics" 2nd Symposium, 391 - 428, Academic Press, New York.

Liepmann, H. W., Narasimha, R., and Chahine, M. T. (1962), Phys. Fluids 5, 1313 - 1324.

Mott - Smith, H. M. (1951), Phy. Rev. 82, 885 - 892.

Muckenfuss, C. (1960), Phy. Fluids 3, 320 - 321.

Rosen, P. (1954), J. App. Phy. 25, 336 - 338.

Sakurai, A. (1957), J. Fluid Mech. 3, 255 - 260.

Shen, S. F. (1963), in "Rarefied Gas Dynamics, Third Symposium" Vol. II, 112 - 131, Academic Press, New York.

Sherman, F. S., and Talbot, L. (1960) in "Rarefied Gas Dynamics", 1st Symposium, 161 - 191, Pergamon Press, London.

Tsien, H. S. (1946), J. Aero. Sci. 13, 653 - 664.

Willis, D. R. (1958), Princeton Univ. Aero. Eng. Dept. Rept. 442.

Willis, D. R. (1962), Phys. Fluids 5, 127 - 135.

Zoller, K., Tech. Note BN - 83, (1951), Zest. Phy. 130, 1 - 38.

BASIC FLUID DYNAMICS

by

S. F. Shen
Cornell University

I. Introduction

In attempting to survey basic fluid dynamics in a program dedicated to the field of applied mathematics in space problems, the foremost question is to settle upon what should be meant by the word "basic". To this end, Professor Goldstein's admirable monograph (S. Goldstein, "Lectures in Fluid Dynamics", Interscience Publishers, 1960) has provided a valuable guiding principle. Our endeavor in the following, however, is slightly different from an abbreviated version of Goldstein's book, but reflects somewhat the aerodynamicist's viewpoint. After the formulation of the general equations of motion, the emphasis is mostly on the motivation and derivation of the different approximations which find applications in various practical problems, particularly to bodies in flight at the higher speed ranges typical of space activities. Much material of basic and mathematical interest is unavoidably left out, as are the full details of the solution of any specific problem. In their places, we choose rather to illustrate, ever so briefly to be sure, how the theory has been exploited in the explanation and prediction of complicated physical phenomena.

Since most of the coverage is "basic", therefore contained in the well-known treatises such as those of Lamb and Milne-Thomson, as well as Goldstein's book mentioned above, we have refrained from giving references except in rare instances.

II. Description of Fluid Motion

The fluid medium we work with shall be a continuum which, although somewhat idealized, should approximate the real gas of interest, namely air, in its behavior. Fluid dynamics then deals with such a gas in motion with or without the presence of solid boundaries. The state of gas in equilibrium, as when enclosed in a stationary and insulated vessel, is described by two thermodynamic variables, say density ρ and temperature T; and any other thermodynamic variable can be expressed in terms of ρ and T. In particular, we often desire to know the pressure p of the gas, observable as the normal force per unit area acting on the wall. The relation may be written as

$$p = p(\rho, T) \qquad (II.1)$$

and usually referred to as the "equation of state". Under the assumption of a perfect gas, Eq. (II.1) becomes explicitly

$$p = \rho R T \qquad (II.2)$$

where R is the gas constant, depending only upon the molecular weight of the gas.

When a body of gas is in arbitrary motion, it becomes necessary to regard the body of gas as composed of a large number of fluid elements, which must be small enough to represent the details of the fluid motion, yet not so small as to exhibit the coarse nature of the molecular motion. A velocity $\vec{V}$ may be assigned to each fluid element, and an observer riding with the fluid element may now determine the density and temperature of the gas in the fluid element. The pressure p follows again from Eq. (II.1). If we trace the changes of $p, \rho, T, \vec{V}$ with time for each fluid element, the result is the "Lagrangian description" of the fluid motion. Alternatively, it is often more convenient for analysis to use a field representation by examining the flow pattern, i.e., the functions

$$p(\vec{r}, t) \ , \ \rho(\vec{r}, t) \ , \ T(\vec{r}, t) \ , \ \vec{V}(\vec{r}, t)$$

where $\vec{r}$ designates the location of the fluid element at the given time t. This is now the "Eulerian description" of the fluid motion.

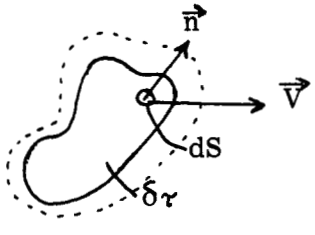
In the Eulerian description, the rate of change of any property Q of a <u>given</u> fluid element is usually written as DQ/Dt. Hence if Q is expressible as $Q(\vec{r}, t)$, we have

$$\begin{aligned} \frac{D}{Dt} Q &= \lim_{\Delta t \to 0} \left[Q(\vec{r}+\Delta\vec{r}, t+\Delta t) - Q(\vec{r}, t) \right] / \Delta t \\ &= \lim_{\Delta t \to 0} \left[Q(\vec{r}+\vec{V}\Delta t, t+\Delta t) - Q(\vec{r}, t) \right] / \Delta t \\ &= \left(\frac{\partial}{\partial t} + \vec{V} \cdot \nabla \right) Q . \end{aligned} \tag{II.3}$$

For example for given $\vec{V}(\vec{r}, t)$, the acceleration of the fluid element is equal to $D\vec{V}/Dt$. However, sometimes Q may not be given as a field, then a direct evaluation is necessary. To illustrate the latter, let Q be the volume $\delta\tau$ of a fluid element, and define the "dilatation" θ as the rate of volume change of the fluid element, per unit volume:

$$\theta \equiv \lim_{\delta\tau \to 0} \frac{1}{\delta\tau} \left(\frac{D}{Dt} \delta\tau \right). \tag{II.4}$$

If $\delta\tau$ is bounded by surface S and $\vec{n}$ is the outward unit normal on the surface element dS, clearly by definition



$$\begin{aligned} \theta &\equiv \lim_{\delta\tau \to 0} \int_S \vec{V} \cdot \vec{n} \, dS / \delta\tau \\ &= \operatorname{div} \vec{V} \ . \end{aligned} \tag{II.5}$$

Of considerable interest in the Eulerian description of fluid motion is the "streamline pattern", showing the direction of motion of each fluid element at a given instant. The "streamlines" are defined by

$$d\vec{r}_s \times \vec{V} = 0, \qquad (II.6)$$

where $d\vec{r}_s$ is a length element on the streamline. If the flow pattern does not vary with time, the fluid motion is said to be a "steady flow". The streamlines in such cases coincide with the trajectories of the fluid elements.

Aside from the translational motion of the fluid element, we must, of course, also expect in general a rotational motion as well as a change of shape with time. The angular velocity of the fluid element turns out to be one half the "vorticity" $\vec{\omega}$, which is defined through a given velocity field $\vec{V}(\vec{r}, t)$ as

$$\vec{\omega} \equiv \nabla \times \vec{V} \quad . \qquad (II.7)$$

Following Eq. (II.6), we may then look at the vorticity pattern by introducing "vortex lines" analogous to the streamlines,

$$d\vec{r}_v \times \vec{\omega} = 0 \qquad (II.8)$$

where $d\vec{r}_v$ is a length element on the vortex line.

III. Equations of Fluid Motion

The equations of fluid motion express the requirements that the fundamental laws of the conservation of mass, momentum and energy must not be violated. These can be very simply stated if the Lagrangian description is adopted. Consider a small fluid element of volume $\delta\tau$; its mass will be $\rho\delta\tau$. For generality we introduce a "mass source" $\dot{m}$ such that mass is being added to the fluid element at the rate of $\dot{m}\delta\tau$. Then the law of conservation of mass as applied to $\delta\tau$ states that

$$\frac{D}{Dt}\rho\delta\tau = \dot{m}\delta\tau \; . \qquad (III.1)$$

In Lagrangian sense, the left hand side is an ordinary time derivative of a product, and we may write

$$\delta\tau \frac{D\rho}{Dt} + \rho \frac{D}{Dt}\delta\tau = \dot{m}\delta\tau \; .$$

By using the definition of the dilitation θ, Eq. (II.4), to evaluate $\frac{D}{Dt}\delta\tau$, the result may be rewritten as

$$\frac{D\rho}{Dt} + \rho\theta = \dot{m} \qquad (III.2)$$

which is known as the "equation of continuity". When ρ, $\vec{V}$ and $\dot{m}$ are regarded as

field quantities in Eulerian description, we only need to interpret the terms in Eq. (III.2) according to Eqs. (II.3) and (II.5).

For the momentum $\rho\vec{V}\delta\tau$ and energy $\rho E\delta\tau$ (E being defined as the energy per unit mass of the fluid), equations similar to Eq. (III.1) may be written with a "momentum source" $\vec{P}$ and an "energy source" $\dot{E}$, respectively, on the right hand side. The same manipulation yields

$$\rho\frac{D\vec{V}}{Dt} = \vec{P} - \dot{m}\vec{V} \qquad (III.3)$$

and

$$\rho\frac{DE}{Dt} = \dot{E} - \dot{m}E \qquad (III.4)$$

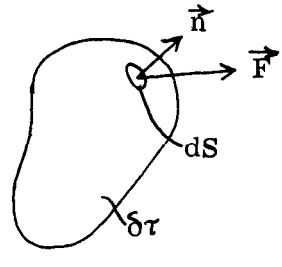
Again, although derived from the Lagrangian description, Eqs. (III.3) and (III.4) offer no difficulty in interpretation for the Eulerian description, provided $\dot{m}$, $\vec{P}$ and $\dot{E}$ are given as field quantities.

We restrict ourselves in the following to the case of $\dot{m} = 0$. To proceed further with the terms $\vec{P}$ and $\dot{E}$, it will be assumed that these are only due to the interactions between adjacent fluid elements, and that the basic fluid properties are isotropic, namely, invariant with orientation. For $\vec{P}$, aside from the pressure p experience shows that any non-uniformity of motion causing a change of shape of the fluid element would be resisted by the fluid through the development of internal stresses between fluid elements. For $\dot{E}$, experience shows that heat will flow through the boundary of the fluid element if a non-uniformity of temperature exists. In addition, the stresses acting on the boundary perform mechanical work on the fluid element.

Consider now a fluid element $\delta\tau$ within the surface S. On a surface element dS, let $\vec{n}$ be the unit outward normal and $\vec{F}$ the resultant stress vector. Referring to a set of Cartesian coordinates x_i, $i = 1,2,3$, these have components n_i and F_i, resp. It is then convenient to introduce a stress tensor τ_{ij} such that

$$F_i = \tau_{ij}n_j \,. \qquad (III.5)$$

In Eq. (III.5) and hereafter, the customary convention of summing over an identical subscript will be understood. Since τ_{ij} includes the pressure which is present even without fluid motion, we may separate τ_{ij} into two parts:



$$\tau_{ij} = \tau'_{ij} - p\delta_{ij}$$

$$(\delta_{ij} = 0,\ i \neq j\ ;\ \delta_{ij} = 1,\ i = j), \qquad \text{(III.6)}$$

the negative sign indicating that the pressure is always opposite to $\vec{n}$. The tensor τ'_{ij} is the "viscous part" of τ_{ij}, and remains to be related to the non-uniformity of the fluid motion.

If the non-uniformity of the fluid motion is slight, it may be characterized by the first derivatives of $\vec{V}$ with respect to the space variables at the point under consideration, hence the tensor $\partial u_i / \partial x_j$. Splitting $\partial u_i / \partial x_j$ into symmetrical and anti-symmetrical parts, we have

$$\frac{\partial u_i}{\partial x_j} = e_{ij} + \omega_{ij}$$

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \qquad \text{(III.7)}$$

$$\omega_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right),$$

The anti-symmetrical part ω_{ij} is easily seen to be

$$\omega_{ij} = -2\epsilon_{ijk}\omega_k \qquad \text{(III.8)}$$

where ω_k is the component of the vorticity $\vec{\omega}$ defined by Eq. (II.7) and ϵ_{ijk} is the alternating symbol,

$\epsilon_{ijk} = 0$ when the subscripts are not all different;

$= 1$ when i, j, k follow the cyclic order $1, 2, 3$;

$= -1$ when i, j, k do not follow the cyclic order $1, 2, 3$.

Thus ω_{ij} represents the nonuniformity due to a rigid rotation of the fluid. The change of shape of the fluid element as it moves along is entirely represented by the symmetrical tensor e_{ij}. To proceed further, the "viscous hypothesis" is made that τ'_{ij} should be linearly proportional to e_{ij}, i.e.,

$$\tau'_{ij} = C_{ijk\ell} e_{k\ell} .$$

$C_{ijk\ell}$ being constants. Now the physical law must not be affected by the orientation of x_1, x_2, x_3 in an isotropic fluid. Then there must be *

$$C_{ijk\ell} = \lambda\delta_{ij}\delta_{k\ell} + \mu(\delta_{ik}\delta_{j\ell} + \delta_{jk}\delta_{i\ell})$$

reducing to two constants λ and μ. Hence

* See, E.g., Jeffrey: "Cartesian Tensor", p. 70, Cambridge Press, 1931.

$$\tau'_{ij} = \lambda\delta_{ij} e_{kk} + 2\mu e_{ij} \qquad \text{(III.9)}$$

where obviously $e_{kk} = \nabla \cdot \bar{V}$.

There are then normal viscous stresses τ'_{11}, τ'_{22}, τ'_{33}. Summing the three, we have

$$\tau'_{ii} = (3\lambda + 2\mu)\, e_{ii} \ .$$

Thus, like pressure p, the average of the normal viscous stresses is independent of the axes. It is however proportional to the dilitation, and the coefficient $3\lambda + 2\mu$ is referred to as the "bulk viscosity coefficient". For monatomic gases, kinetic theory shows that

$$3\lambda + 2\mu = 0 \quad \text{or} \quad \lambda = -\frac{2}{3}\mu \ .$$

This result is generally assumed in most applications involving air (even though it is composed of primarily diatomic gases) so that the viscous stresses are all proportional to a single material constant, the "viscosity coefficient". Eq. (III.9) becomes

$$\tau'_{ij} = -\frac{2}{3}\mu\,\delta_{ij} e_{kk} + 2\mu e_{ij} \qquad \text{(III.9)'}$$

known as the "Navier-Stokes relation". It may be noted here that the viscosity coefficient is mainly a function of the temperature T.

We next turn to the heat flux due to the non-uniformity of the temperature field. Since T is a scalar, the non-uniformity is characterized by a vector ∇T. If $\vec{q}$ is the heat flux vector (the rate of heat flow per unit area), the assumption of linear dependence leads in an analogous manner to

$$\vec{q} = -k\nabla T \qquad \text{(III.10)}$$

where the proportionality constant k is the "coefficient of thermal conductivity". Eq. (III.10) is known as the "Fourier law". From a molecular viewpoint, both μ and k owe their origin to the random motion of the molecules, and these two are closely related. For example, for monatomic gases, kinetic theory predicts

$$k/\mu\, C_v = 5/2$$

where C_v is the specific heat at constant volume.

With Eqs. (III.9)' and (III.10), it is now possible to represent $\vec{P}$ and $\dot{E}$ explicitly. For $\vec{P}$, there is

$$\begin{aligned} P_i \delta\tau &= \int_S F_i \, dS \\ &= \int_S \tau_{ij} n_j \, dS \\ &= \int_{\delta\tau} \frac{\partial}{\partial x_j} \tau_{ij} \, d\tau \ , \text{ by Gauss' theorem; or, as } \delta\tau \to 0 \ , \end{aligned}$$

$$P_i = \frac{\partial}{\partial x_j}\tau_{ij} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\tau'_{ij} . \qquad (III.11)$$

For $\dot{E}$, there is

$$\dot{E}\,\delta\tau = \int_S \vec{F}\cdot\vec{V}\,dS + \int_S \vec{n}\cdot k\nabla T\,dS$$

$$= \int_S (\tau_{ij}u_i + k\frac{\partial T}{\partial x_j})\,n_j\,dS$$

$$= \int_{\delta\tau}\frac{\partial}{\partial x_j}[\tau_{ij}u_i + k\frac{\partial T}{\partial x_j}]\,d\tau$$

where u_i is the component of $\vec{V}$ in x_i-direction. As $\delta\tau \to 0$, it follows

$$\dot{E} = \frac{\partial}{\partial x_j}(\tau_{ij}u_i) + \frac{\partial}{\partial x_j}(k\frac{\partial T}{\partial x_j}) . \qquad (III.12)$$

Hence, with Eq. (III.11) the momentum equation, Eq. (III.3), becomes finally

$$\rho\frac{Du_i}{Dt} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\tau'_{ij} . \qquad (III.13)$$

In the energy equation, Eq. (III.4), we note that for the fluid element in motion,

$$E = U + \frac{1}{2}V^2$$

where U is the internal energy of the fluid element. Together with Eq. (III.12), Eq. (III.4) after simple manipulation becomes finally

$$\rho\frac{DU}{DT} = -p\,e_{jj} + \tau'_{ij}e_{ij} + \frac{\partial}{\partial x_j}(k\frac{\partial}{\partial x_j}) . \qquad (III.14)$$

The second term of the right hand side clearly represents the work done by the viscous stresses, and often is defined as the "dissipation function" Φ. It may be easily verified that $\Phi \geq 0$ when τ'_{ij} is given by Eq. (III.9)'.

Alternative forms of the energy equation, Eq. (III.14), are sometimes useful. For instance, in terms of the entropy S, since by thermodynamic definition

$$T\,dS = dU + p\,d(\frac{1}{\rho})$$

the continuity equation, Eq. (III.2) (with $\dot{m} = 0$), and Eq. (III.14) combined leads to

$$\rho T\frac{DS}{Dt} = \Phi + \frac{\partial}{\partial x_j}(k\frac{\partial T}{\partial x_j}) . \qquad (III.14)'$$

In terms of the enthalpy h, since by thermodynamic definition

$$h = U + \frac{p}{\rho}$$

Eq. (III.14) may also be replaced by

$$\rho \frac{Dh}{Dt} = \frac{Dp}{Dt} + \Phi + \frac{\partial}{\partial x_j}\left(k \frac{\partial T}{\partial x_j}\right). \qquad \text{(III.14)''}$$

IV. Physical Boundary Conditions of Fluid Motion

The fluid motion has been defined in the above through the unknowns p, ρ, T and $\vec{V}$, which are required to satisfy Eqs. (II.2), (III.2) (with $\dot{m} = 0$), (III.13) and (III.14). A typical problem is to find the solution when an obstacle moves through the fluid in a prescribed manner. In the fluid domain, there remains the question of relating the values of these unknowns for the fluid elements in contact with the obstacle, with the prescribed motion and properties of the obstacle itself. We may think of Eqs. (II.2) and (III.2) as defining p and ρ in terms of $\vec{V}$ and T, so Eqs. (III.13) and (III.14) are really the equation to be integrated. Thus, if the obstacle is impermeable and represented by the surface $F_S(x_i, t) = 0$ and its temperature by the condition $T = T_S(t)$ on $F_S = 0$, we are interested to assign values of $\vec{V}$ and T for the fluid elements satisfying $F_S = 0$.

Now the resultant velocity of a point on the obstacle must satisfy $DF_S/DT = 0$. Since the obstacle is assumed to be impermeable, the velocity of the fluid element at the same point must have the same velocity component normal to the surface, and therefore satisfy also $DF_S/Dt = 0$, although the tangential velocity is still arbitrary. We refer to this as the "condition of no penetration", or

$$\frac{DF_S}{Dt} = 0 \quad \text{for fluid elements on } F_S = 0. \qquad \text{(IV.1)}$$

Obviously by the same reasoning, Eq. (IV.1) is also the condition at the interface between two dissimilar fluids.

As for the tangential component of the fluid velocity and the temperature of the fluid element at the boundary, one usually appeals to experience whenever the mathematical solution requires these data. Ordinarily it is assumed that the fluid element shall have neither a relative velocity with respect to the boundary -- the "condition of no slip", nor any temperature differences from that of the boundary -- the "condition of no jump". These are confirmed as first approximations by kinetic theory considerations, so long as the gas is not too rarified.

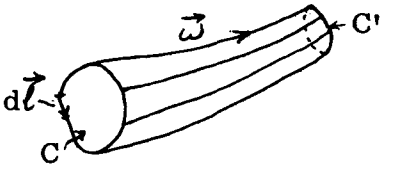
The precise conditions under which the mathematical problem will be "properly set" is in general a difficult question because of the complicated non-linear nature of the equations. The practice is rather to look for a solution when physically the problem is well-defined and can be set up in an experimental investigation. However, it should be noted that empirically under seemingly identical conditions the observed flow may be either "laminar" or "turbulent". Take the steady flow through a circular pipe as an example: A mathematical solution of the equations predicts that the flow should move in layers, and is indeed well confirmed experimentally but only if the flow velocity is relatively small. At higher velocities, the actual flow is composed of a steady mean motion superposed by random time-dependent fluctuations. This phenomenon is typical rather than exceptional. It strongly suggests that the general uniqueness condition for the system of equations describing fluid flow would be extremely difficult to lay down.

V. Rotational and Irrotational Motions

We defined in Eq. (II.7) the vorticity vector $\vec{\omega}$

$$\vec{\omega} = \nabla \times \vec{V}$$

representing the rigid body rotation of the fluid element. Kinematically, the fluid motion may be classified as rotational or irrotational depending on whether $\vec{\omega} \neq 0$ in general or $\vec{\omega} = 0$ everywhere in the fluid. We note first that because $\vec{\omega}$ is the curl of $\vec{V}$, it is a solenoidal vector, i.e., $\nabla \cdot \vec{\omega} = 0$. Now if we have a steady flow field with $\dot{m} = 0$, the equation of continuity, Eq. (III.2), reduces to $\nabla \cdot \rho \vec{V} = 0$. Thus by analogy it can be said that $\vec{\omega}$ also satisfies an "equation of continuity". Let a contour C enclosing a surface S be drawn in the fluid. Vortex lines can be passed through points of C to form a "vortex tube", and the following must hold:



$$\int_S \omega_i \, dS_i = \int_{S'} \omega_i \, dS_i$$

where S' is the area enclosed by C' anywhere downstream along the vortex lines from C. In particular, by taking $S \to 0$, the vortex tube becomes a very thin

"vortex element". Thus a vortex element can never end within the fluid. It may, however, form a closed loop.

If the fluid motion is an irrotational motion, by applying Stokes' theorem to the contour C,

$$\oint_C \vec{V} \cdot d\vec{\ell} = \int_S \omega_i \, dS_i = 0 .$$

Consequently a "velocity potential" ϕ exists such that

$$\vec{V} = \nabla\phi \quad .$$

Since a scalar ϕ defines the velocity vector, the mathematical problem is then to find the solution for a single function ϕ and becomes much simplified. It is therefore of interest to examine the circumstances under which the irrotational approximation may be adopted.

With $\dot{m} = 0$, consider the momentum equation, Eq. (III.3),

$$\frac{D\vec{V}}{Dt} = \frac{1}{\rho}\vec{P} .$$

Expanding $D\vec{V}/Dt$, we have

$$\frac{D\vec{V}}{Dt} = \frac{\partial\vec{V}}{\partial t} + \frac{1}{2}\nabla\vec{V}^2 - \vec{V}\times\vec{\omega} . \qquad (V.1)$$

Hence, by straightforward manipulation and with Eq. (III.2),

$$\nabla \times \frac{D\vec{V}}{Dt} = \frac{\partial}{\partial t}\nabla\times\vec{V} - \nabla\times(\vec{V}\times\vec{\omega})$$

$$= \frac{D\vec{\omega}}{Dt} - (\vec{\omega}\cdot\nabla)\vec{V} - \frac{\vec{\omega}}{\rho}\frac{D\rho}{Dt}$$

$$= \rho\left[\frac{D}{Dt}\left(\frac{\vec{\omega}}{\rho}\right) - \left(\frac{\vec{\omega}}{\rho}\cdot\nabla\right)\vec{V}\right] .$$

The "vorticity equation" follows immediately

$$\frac{D}{Dt}\left(\frac{\vec{\omega}}{\rho}\right) = \left(\frac{\vec{\omega}}{\rho}\cdot\nabla\right)\vec{V} + \frac{1}{\rho}\nabla\times\left(\frac{1}{\rho}\vec{P}\right) . \qquad (V.2)$$

We shall examine the behavior of $\vec{\omega}$ under the following simplifying conditions:

1) $p = p(\rho)$, e.g., $p \propto \rho^{\gamma}$ for isentropic process (γ being the ratio of specific heats), or ρ = const. for incompressible fluid;

2) $\mu = 0$, the inviscid approximation. Under the simplification, Eq. (III.11) gives $\vec{P} = -\nabla p$, and since $\nabla\times\left(\frac{1}{\rho}\nabla p\right) = \nabla\times\left(\nabla\int\frac{dp}{\rho}\right) = 0$, Eq. (V.2) becomes

$$\frac{D}{Dt}\frac{\vec{\omega}}{\rho} = \left(\frac{\vec{\omega}}{\rho}\cdot\nabla\right)\vec{V}$$

or

$$\frac{D}{Dt}\frac{\omega_i}{\rho} = \frac{\omega_j}{\rho}\frac{\partial}{\partial x_j}u_i$$

$$= \frac{\omega_j}{\rho}\left[e_{ij} + \omega_{ij}\right] .$$

Noting Eq. (III.8),

$$\omega_j \omega_{ij} = -2 \epsilon_{ijk} \omega_j \omega_k = 0 \ .$$

We finally get

$$\frac{D}{Dt} \frac{\omega_i}{\rho} = \frac{\omega_j}{\rho} e_{ij} \qquad (V.3)$$

which is sometines interpreted as saying that following the fluid element, ω_i/ρ changes due to the "stretching" of the vortices. In particular, for two-dimensional motion $\vec{\omega} = (0, 0, \omega_3)$ but $e_{33} = 0$, Eq. (V.3) degenerates into

$$\frac{D}{Dt} \frac{\omega_3}{\rho} = 0 \qquad (V.4)$$

saying that the vorticity, strictly speaking ω_3/ρ, is attached to the fluid element without change. Following Eq. (V.4), as long as $p = p(\rho)$ and in the inviscid limit, if at some time the fluid element does not possess vorticity it will not acquire vorticity in two-dimensional motion. When the flow field is set up from rest through the arbitrary movements of a two-dimensional body, we therefore expect irrotational motion at all times. For the general three-dimensional flow, the same conclusion can be reached by integrating Eq. (V.3) for a given fluid element.* These are of course only useful in practical cases when the underlying assumptions are acceptable.

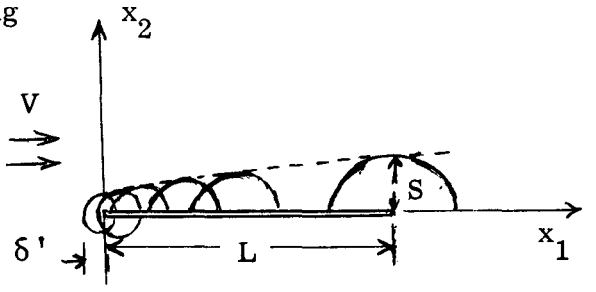
Let us now examine the role of viscosity. Consider for simplicity the small perturbation from a state of rest, i.e., $\vec{V} = \vec{V}'$, $\vec{\omega} = \vec{\omega}'$, $\rho = \rho_0 + \rho'$, etc., ρ_0 being the density of the fluid at rest and primed quantities being the small perturbations. After neglecting the quadratic terms involving the perturbation quantities and with the help of Eqs. (III.11) and (III.9)', Eq. (V.2) is reduced to the diffusion equation

$$\frac{\partial}{\partial t} \vec{\omega}' = \nu_0 \nabla^2 \vec{\omega}' \qquad (V.5)$$

where $\nu_0 = \mu_0/\rho_0$, the kinematic viscosity. If a vortex element is generated in an infinite fluid at $t = 0$ and maintained afterwards, the consequence of Eq. (V.5) is that the vorticity will spread out, with decreasing strength, to occupy a region of size $(\nu_0 t)^{1/2}$ beyond which the effect is essentially nil. This result is qualitatively

* See e.g., L. M. Milne-Thomson: Theoretical Hydrodynamics, 4th ed., Macmillan (1960), p. 84.

useful in visualizing the flow patterns surrounding a body moving in a fluid at rest. Suppose a thin two-dimensional plate of length L moves parallel to itself in a viscous fluid at constant velocity V. An obvious irrotational solution is that the fluid is undisturbed, satisfying all differential equations except for the viscous "no-slip" and "no-jump" conditions at the surface. We imagine viscosity to be absent for $t < 0$, but suddenly turned on at $t = 0$. The fluid elements in contact with the plate will be instantaneously arrested, creating a surface of discontinuity which may be interpreted as a vortex sheet composed of concentrated vortex elements. The vorticity subsequently spreads out approximately at a rate $O(\sqrt{\nu_0 / t})$. To an observer fixed relative to the plate, the instantaneous flow pattern will be swept downstream at a speed equal to V and the vorticity will be seen as essentially confined in a region roughly parabolic starting from the leading edge of the plate, at the end of the plate the thickness δ reaching a value $O(\sqrt{\nu_0 L / V})$. In non-dimensional form, we have therefore



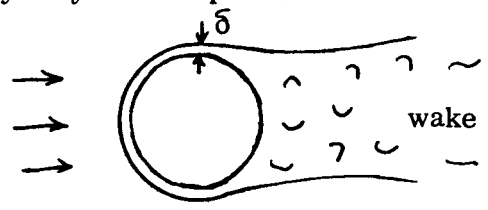
$$\frac{\delta}{L} \sim O(1/\sqrt{Re})$$

where $Re = VL/\nu_0$, the "Reynolds number" based on the length L. There would be furthermore a disturbed region ahead of the plate of size δ', given by

$$\delta' \sim O(\sqrt{\nu_0 \delta'/V})$$

or, again in terms of a Reynolds number, $Re_{\delta'} \equiv V\delta'/\nu_0 \sim O(1)$. Thus the size of the region of rotational flow because of the viscous effects is confined to the immediate neighborhood of the plate as $\nu_0 \to 0$. In fact, the thickness δ' tends to zero much faster than the thickness δ. The layer of $O(\delta)$ adjacent to the body is referred to as the "boundary layer". The viscous and rotational region swept behind the body is the "wake". Outside of the thin boundary layer and the wake, the flow is seen to be essentially irrotational. For blunt bodies these qualitative descriptions remain valid, but although the boundary layer thickness is still proportional to $\sqrt{\nu_0}$, the wake will be of the order of the body

thickness. It should be mentioned that rotationality may also be present due to curved shock waves which form ahead of the body when it moves at high speeds. (See §X). This is an example where $p = p(\rho)$ is not true.



Finally we note that the boundary layer and the wake are actually the corrections to an inviscid solution due to the viscous boundary conditions. Consequently outside of these regions whether the flow be rotational or irrotational, the fluid may be regarded as inviscid. As the kinematic viscosity of gases is usually very small, in most flow problems the Reynolds number will be large and the boundary layer will be relatively thin. Then the inviscid "no penetration" condition may be applied without serious error as if the boundary layer were absent. The difficulty of the unknown boundary of the wake, however, cannot be circumvented in constructing an inviscid approximation for blunt bodies.

VI. The Inviscid Approximation

Let us now exploit the inviscid approximation. Since the viscosity μ and the thermal conductivity k are of the same mechanism, the fluid should also be regarded as non heat-conducting in the same approximation. The immediate consequence from Eq. (III.14)' is

$$\frac{DS}{Dt} = 0 \qquad \text{(VI.1)}$$

i.e., the entropy is constant following each fluid element, though not necessarily throughout the flow field. The "Navier-Stokes' equations", (Eq. III.13), degenerate into the "Euler equations"

$$\rho \frac{D\vec{V}}{Dt} = -\nabla p. \qquad \text{(VI.2)}$$

The continuity equation, Eq. (III.2), of course is unaffected:

$$\frac{D\rho}{Dt} + \rho \nabla \cdot \vec{V} = 0. \qquad \text{(VI.3)}$$

Consider again a small perturbation of the fluid from rest at pressure p_0 and density ρ_0. Neglecting quadratic terms of the perturbation quantities $\vec{V}'$, p' and

ρ', we get the "acoustic theory" from Eqs. (VI.1-3):

$$\left.\begin{aligned} S &= S_0, \text{ const.} \\ \rho_0 \frac{\partial \vec{V}}{\partial t} &= -\nabla p' \\ \frac{1}{\rho_0}\frac{\partial \rho'}{\partial t} &+ \nabla \cdot \vec{V}' = 0. \end{aligned}\right\} \qquad \text{(VI.4)}$$

The first of these may alternatively be expressed as

$$p/p_0 = (\rho/\rho_0)^{\gamma}, \quad \gamma = C_p/C_v,$$

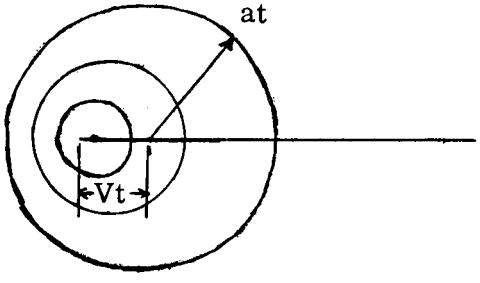
or

$$p' = a^2 \rho' \qquad \text{(VI.5)}$$

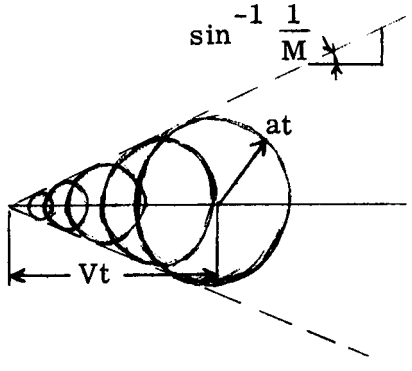
where $a = \sqrt{(\partial p/\partial \rho)_S} = \sqrt{\gamma p_0/\rho_0}$, the "speed of sound". Eliminating p' and $\vec{V}'$ in favor of ρ', we find

$$\frac{\partial^2}{\partial t^2}\rho' - a^2 \nabla^2 \rho' = 0 \qquad \text{(VI.6)}$$

which is the "wave equation". The elementary solution for introducing a small disturbance at a point at $t = 0$ is such that the disturbance spreads out in space at a rate equal to the sound speed "a", and beyond a radius of at the fluid is undisturbed. For a source of disturbance moving at constant velocity $\vec{V}$, to an observer fixed to the source of disturbance, two different flow patterns result depending on the "Mach number" $M \equiv V/a$. For $M < 1$ the disturbance spreads out in all directions, eventually swallowing up the entire space in a long enough time. For $M > 1$, the disturbed region is conical, formed by the envelope to the drifting spheres, with the



$M < 1$



$M > 1$

vertex at the source of disturbance, the semi-angle being equal to the "Mach angle" $\sin^{-1}\frac{1}{M}$. The conical surface itself for finite time, in conjunction with the spherical

surface in the back, is the wave front separating the undisturbed and the disturbed regions. Following von Kármán, one may refer to the undisturbed region ahead of the conical surface as the "zone of silence", and the disturbed region behind as the "zone of action". While the above is based upon linearized small perturbation theory, the difference in behavior of subsonic and supersonic flows remains qualitatively the same even if the disturbances caused by the moving object are no longer small.

Without restricting ourselves to small disturbances, we return to Eqs. (VI.1) to (VI.3). In certain cases, a first integral of the Euler equations, Eq. (VI.2), can be directly obtained, and as a result further simplify the problem of finding a solution. By Eq. (V.1), Eq. (VI.2) may be written as

$$\frac{\partial \vec{V}}{\partial t} + \frac{1}{2}\nabla \vec{V}^2 - \vec{V}\times\vec{\omega} = -\frac{1}{\rho}\nabla p \ .$$

Now the definition of entropy S is, for a given element, $T\,dS = dh - \frac{1}{\rho}dp$. But inasmuch as T and ρ are always expressible as functions of p and h, this expression may also be regarded as an ordinary differential relation defining $S(p,h)$, hence leading to

$$T\nabla S = \nabla h - \frac{1}{\rho}\nabla p \ .$$

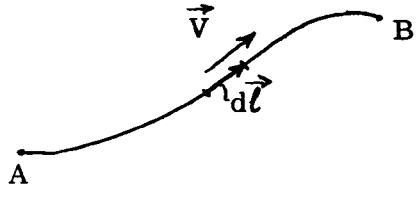
We define next a "stagnation enthalpy" H,

$$H \equiv h + \frac{1}{2}\vec{V}^2$$

and recast Eq. (VI.2) into

$$\frac{\partial}{\partial t}\vec{V} + \nabla H - T\nabla S = \vec{V}\times\vec{\omega}\,. \qquad (VI.7)$$

For the special case of steady flow, $\frac{\partial}{\partial t}\vec{V} = 0$, Eq. (VI.7) is known as "Crocco's theorem", showing for instance that vorticity would arise due to entropy gradient. If Eq. (VI.7) is dotted into the length element $d\vec{\ell}$ along a streamline, at given t, and then integrated between the end points A and B, the result is



$$\frac{\partial}{\partial t}\int_A^B \vec{V}\cdot d\vec{\ell} + \int_A^B dH - \int_A^B T\,dS = 0.$$

This yields the so-called "Bernoulli's equation" in the following cases:

a) For steady flow $\left(\frac{\partial}{\partial t} = 0\right)$, $\frac{DS}{Dt} = \vec{V}\cdot\nabla S = 0$, hence $\vec{V}$ and $d\vec{\ell}$

are both normal to ∇S. Consequently

$$H = \text{const. along any streamline} \qquad (VI.8)$$

b) For irrotational flow with uniform entropy everywhere ("homentropic"). $\vec{V} = \nabla\phi$ and $\nabla S = 0$, hence

$$\frac{\partial\phi}{\partial t} + H = \text{const. along any streamline.} \qquad (VI.9)$$

When the streamlines can always be traced to a region of steady uniform flow, the constant in the right-hand side of Eq. (VI.8) or (VI.9) becomes identical for all points in the flow field.

We next proceed to derive the equation for the velocity potential ϕ in an irrotational homentropic flow. The scalar product of $\vec{V}$ with Eq. (IV.2) leads to

$$\frac{\partial}{\partial t}\frac{V^2}{2} + \frac{1}{2}(\vec{V}\cdot\nabla)V^2 = -\frac{a^2}{\rho}\vec{V}\cdot\nabla\rho$$

$$= \frac{a^2}{\rho}\left[\frac{\partial\rho}{\partial t} + \rho\nabla\cdot\vec{V}\right],$$

by Eq. (VI.3). But differentiation of Eq. (VI.9) gives

$$\frac{a^2}{\rho}\frac{\partial\rho}{\partial t} + \frac{\partial}{\partial t}\frac{V^2}{2} + \frac{\partial^2\phi}{\partial t^2} = 0 .$$

Eliminating $\frac{\partial\rho}{\partial t}$ between the two expressions, we get

$$\frac{\partial}{\partial t}V^2 + \frac{\partial^2\phi}{\partial t^2} - a^2\nabla\cdot\vec{V} + \vec{V}\cdot\nabla\frac{V^2}{2} = 0 . \qquad (VI.10)$$

In Cartesian coordinates, Eq. (VI.10) may be written as

$$\phi_{tt} - (a^2 - \phi_x^2)\phi_{xx} - (a^2 - \phi_y^2)\phi_{yy} - (a^2 - \phi_z^2)\phi_{zz}$$

$$+ 2(\phi_x\phi_y\phi_{xy} + \phi_x\phi_z\phi_{xz} + \phi_y\phi_z\phi_{yz}$$

$$+ \phi_x\phi_{xt} + \phi_y\phi_{yt} + \phi_z\phi_{zt}) = 0. \qquad (VI.10)'$$

Here a^2 is expressible also in ϕ by noting that

$$H = \frac{\gamma}{\gamma-1}\frac{p}{\rho} + \frac{V^2}{2} = \frac{a^2}{\gamma-1} + \frac{V^2}{2} \qquad (VI.11)$$

while Eq. (VI.9) shows that H is directly related to $\partial\phi/\partial t$. It is, however, more instructive without explicitly evaluating a^2. To fix ideas, suppose we have a body of characteristic length L, characteristic velocity V_∞ in an unsteady motion of

characteristic time t_∞. We assume that generally $a \sim O(a_\infty)$, a_∞ being the characteristic sound speed. Then in Eq. (VI.10)' appear the dimensionless parameters

Mach no. $M_\infty \equiv V_\infty / a_\infty$

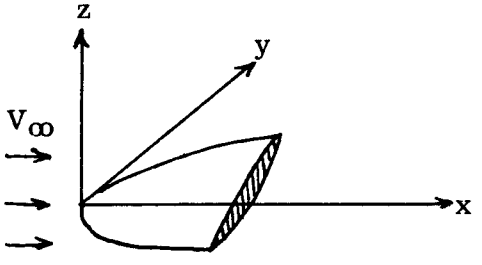
and Strouhal no. $x \equiv L/V_\infty t_\infty$. If $x \sim O(1)$, as $M_\infty^2 \to 0$ the equation reduces to the Laplace equation

$$\nabla^2 \phi = 0 \ . \qquad \text{(VI.12)}$$

Eq. (VI.12) constitutes the (inviscid) "incompressible approximation" since the equation can also be directly obtained by setting $D\rho/Dt = 0$ in the equation of continuity and then using $\vec{V} = \nabla\phi$. The velocity potential now may be solved from the prescribed normal derivative of ϕ on the body surface (to satisfy the condition of "no penetration"). After ϕ is obtained, the Bernoulli's equation, being an integral of the momentum equation, determines the pressure field which simultaneously must co-exist. We shall not discuss the various techniques of solving Laplace's equation.

VII. Small Perturbation Theory for the Steady Flow over Thin Bodies

To illustrate the behavior of the solution of Eq. (VI.10)' when the incompressible approximation is not applicable, consider a uniform stream of velocity V_∞ in the x-direction flowing over a fixed thin body lying close to the x,y-plane. For sufficiently thin bodies, the uniform stream will only be slightly disturbed, and we put the resultant velocity potential as the superposition of that for the uniform stream and a small perturbation, i.e.,



$$\phi = V_\infty x + \phi' \ , \quad \phi_x', \ \phi_y', \ \phi_z' \ll V_\infty \ .$$

Eq. (VI.11) further for steady case becomes

$$a^2 - a_\infty^2 = \frac{\gamma-1}{2}(V_\infty^2 - V^2) \ . \qquad \text{(VII.1)}$$

After substitution of the above into Eq. (VI.10)' and retaining only linear terms in

ϕ' , we get

$$(1-M_\infty^2)\,\phi'_{xx} + \phi'_{yy} + \phi'_{zz} \cong 0 \qquad \text{(VII.2)}$$

provided $|1-M_\infty^2| \sim O(1)$. By a simple stretching of the coordinates

$$x' = x/\sqrt{|1-M_\infty^2|}\,, \quad y' = y, \quad z' = z,$$

Eq. (VII.2) reduces to

$$\pm\,\phi'_{x'x'} + \phi'_{y'y'} + \phi'_{z'z'} = 0\,,$$

the "$\pm$" corresponds to $M_\infty^2 \lesseqgtr 1$. Thus the subsonic flows all satisfy Laplace's equation while the supersonic flows satisfy the wave equation. In fact, by examining the transformed boundary conditions in the new coordinates, it follows readily that flows over a class of bodies at different Mach numbers can be related to each other. The interpretation of a known flow over a given body and Mach number as that for a different body at a different Mach number is referred to as the "similarity rule". In subsonic flows, such is known as the "Prandtl-Glauert rule", in supersonic flows, the "Ackeret rule".

The linearized theory, Eq. (VII.2), fails when some of the neglected terms become comparable with those retained. If we evaluate the neglected terms, it may be verified that the above linearization implies

(i) $\quad a^2 - (V_\infty + \phi'_x)^2 \sim a_\infty^2 - V_\infty^2 >> V_\infty \phi'_y$ or $V_\infty \phi'_z$

(ii) $\quad a \sim a_\infty >> \phi'_y\,,\ \phi'_z$.

The condition (i) breaks down where $a_\infty^2 \cong V_\infty^2$, or $M_\infty^2 \cong 1$, i.e., in "transonic flows". The condition (ii) breaks down when $a_\infty << V_\infty$, or $M_\infty >> 1$, i.e., in "hypersonic flows". In both transonic and hypersonic cases, then, we are forced to non-linear theories even for small perturbations.

Let us demonstrate briefly the complications of the transonic approximation. If $V_\infty \cong a_\infty$, it is convenient to consider the flow over a thin body as small perturbations on a uniform sonic flow ($V_\infty = a_\infty = a^*$, say) without the body. Thus putting

$$\phi = a^* x + \phi'$$

and rewriting Eq. (VII.1)

$$a^2 = \frac{\gamma+1}{2}\,a^{*2} - \frac{\gamma-1}{2}\,V^2 \cong a^{*2} - \frac{\gamma-1}{2}\,[\,2a^*\phi'_x\,] + \cdots\,,$$

we get from Eq. (VI.10)' after retaining all quadratic terms,

$$-(\gamma+1)\phi'_x\phi'_{xx} + a^*(\phi'_{yy} + \phi'_{zz}) - 2(\phi'_y\phi'_{xy} + \phi'_z\phi'_{xz}) = 0 .$$

The essential features remain unchanged by restricting ourselves to two-dimensional motion in the x, z-plane:

$$-(\gamma+1)\phi'_x\phi'_{xx} + a^*\phi'_{zz} - 2\phi'_z\phi'_{xz} = 0 .$$

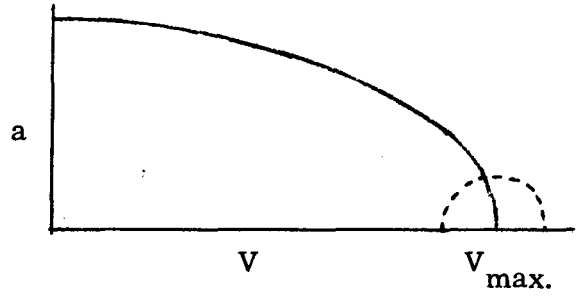
Here one or both of the quadratic terms must be everywhere of the same order as the term $a^*\phi'_{zz}$. In order to do so, clearly the function ϕ' must vary much more rapidly in the x-direction than in the z-direction. Hence the first term should dominate, and the "transonic equation" for two-dimensional steady flow finally reduces to

$$-(\gamma+1)\phi'_x\phi'_{xx} + a^*\phi'_{zz} = 0 . \qquad \text{(VII.3)}$$

The "similar rule" for relating the transonic flows over geometrically similar bodies of different thickness to each other was deduced by von Kármán.

In analogous manner, the non-linear perturbation equation for the velocity potential and the similarity rule in hypersonic flow have been given by Tsien. However, strong curved shocks inevitably occur in hypersonic flow, and the flow behind the shock and over the body is generally rotational. Tsien's equation therefore loses much of its significance. On the other hand, if we plot Eq. (VII.1), in hypersonic flow the sound speed a and the resultant velocity V will always be in the region near the maximum velocity $V_{max.}$.

$$V_{max.} = V_\infty \left[1 + \frac{2}{(\gamma-1)M_\infty^2} \right] .$$



For considerable variation of the local Mach number, the resultant velocity V is essentially unchanged. In addition, the streamlines around thin bodies are always only slightly inclined. Consequently, it is obvious that the perturbation velocity $u' \ll v', w'$. Neglecting u' completely, the steady flow pattern in the y, z-plane at different streamwise stations x can be interpreted as the unsteady flow pattern in the y, z-plane at successive times, the elapsed time Δt between two stations Δx apart being given by $\Delta t \cong \Delta x / V_\infty$.

This is the essence of Hayes' "equivalence principle", which holds regardless of whether the flow is rotational of irrotational, of whether any shock wave occurs at the nose of the body. It simplifies the problem of hypersonic steady flow over a thin body by reducing it to an unsteady flow over a body of lesser dimension.

VIII. One-dimensional Unsteady Flow and the Formation of Shock

We now return to Eq. (VI.10)' but restrict ourselves to one-dimensional unsteady flows,

$$\phi_{tt} - (a^2 - \phi_x^2)\phi_{xx} + 2\phi_x\phi_{xt} = 0 \,; \tag{VIII.1}$$

According to the theory of quasi-linear partial differential equations, this equation is hyperbolic, just as in the acoustic approximation, since the discriminant

$$(2\phi_x)^2 + 4(a^2 - \phi_x^2) = 4a^2 > 0 .$$

Thus there exist real characteristic curves, along which the values of ϕ_x and ϕ_t may be described without uniquely determining the higher derivatives ϕ_{xx}, ϕ_{xt}, ϕ_{tt}. Let the running variable along such a characteristic curve be σ. For prescribed ϕ_x and ϕ_t along the curve, the following must hold

$$\left.\begin{aligned} \phi_{x\sigma} &= \phi_{xx}x_\sigma + \phi_{xt}t_\sigma \\ \phi_{t\sigma} &= \phi_{tx}x_\sigma + \phi_{tt}t_\sigma \end{aligned}\right\} \tag{VIII.2}$$

We normally should be able to solve ϕ_{xx}, ϕ_{xt} and ϕ_{tt} from Eqs. (VIII.1) and (VIII.2), except when x_σ and t_σ are such that the matrix

$$\begin{pmatrix} 1 & 2\phi_x & -(a^2 - \phi_x^2) & 0 \\ t_\sigma & x_\sigma & 0 & \phi_{t\sigma} \\ 0 & t_\sigma & x_\sigma & \phi_{x\sigma} \end{pmatrix}$$

has rank 2. Hence, to require the curve be a characteristic

$$\begin{vmatrix} 1 & 2\phi_x & -(a^2 - \phi_x^2) \\ t_\sigma & x_\sigma & 0 \\ 0 & t_\sigma & x_\sigma \end{vmatrix} = 0$$

or

$$\frac{dx}{dt} \equiv \frac{x_\sigma}{t_\sigma} = \phi_x \pm a \tag{VIII.3}$$

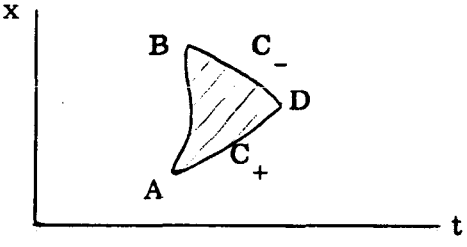
giving the direction of the characteristics. Also

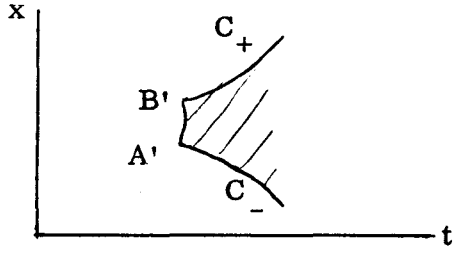
$$\begin{vmatrix} 1 & -(a^2 - \phi_x^2) & 0 \\ t_\sigma & 0 & \phi_{t\sigma} \\ 0 & x_\sigma & \phi_{x\sigma} \end{vmatrix} = 0$$

or $$\phi_{t\sigma} = \phi_{x\sigma}(-\phi_x \pm a), \qquad \text{(VIII.4)}$$

giving a condition on the variation of ϕ_x and ϕ_t along the characteristics. Thus we have two families of characteristics, which may be referred to as the $C_\pm$ curves, resp., according to the sign "$\pm$" in Eqs. (VIII.3) and (VIII.4).

By using the Bernoulli equation, Eq. (VI.9), the sound speed "a" may be related to ϕ_x and ϕ_t. If initial data are prescribed along an ordinary curve (not coincident with either characteristic) in the x,t-plane, it is known that the characteristics relation Eqs. (VIII.3) and (VIII.4) uniquely determine the solution in the curvilinear triangle ABD, bounded by the characteristics C_+ and C_- through A and B, resp. The segment AB is the "domain of dependence" for point D. Likewise, if data are modified along a segment A'B', the solution in the shaded region shown in the sketch will be affected and is the "range of influence" of the segment A'B'. Moreover, the higher order derivatives normal to a characteristic may be discontinuous. Consequently, a characteristic , and a characteristic only, can serve as the boundary between regions of constant state and variable flow, provided Eq. (VIII.1) holds everywhere.





Now, by differentiating the Bernoulli's equation along a characteristic, there follows

$$\phi_{t\sigma} + \phi_x \phi_{x\sigma} + \frac{2a}{\gamma - 1} a_\sigma = 0 \ .$$

Because of Eq. (VIII.4), it reduces to

$$\frac{d}{d\sigma}\left[\pm\phi_x + \frac{2}{\gamma-1}a\right] = 0 .$$

Thus, if we define the "Riemann invariants" r and s as

$$\left.\begin{aligned} r &= \frac{1}{2}\phi_x + \frac{a}{\gamma-1} \\ s &= \frac{1}{2}\phi_x - \frac{a}{\gamma-1} \end{aligned}\right\} \qquad \text{(VIII.5)}$$

it follows that

$$r = r(\alpha), \quad s = s(\beta)$$

where α = const. along the C_+-curves and β = const. along the C_--curves. (The running variable σ becomes β along C_+ and α along C_-.) Equivalently, since by Eq. (VIII.3)

$$x_\beta/t_\beta = \phi_x + a, \quad x_\alpha/t_\alpha = \phi_x - a,$$

we have

$$\left.\begin{aligned} \frac{\partial r}{\partial t} + (a+\phi_x)\frac{\partial r}{\partial x} &= 0 \\ \frac{\partial s}{\partial t} + (-a+\phi_x)\frac{\partial s}{\partial x} &= 0 . \end{aligned}\right\} \qquad \text{(VIII.6)}$$

The property r is thus propagated forward without change at the local sound speed relative to the fluid, while the property s is propagated backward without change at the local sound speed relative to the fluid.

It is clear that in general a region in the x,t-plane may be mapped to a region in the r,s-plane through one-to-one correspondence. However, there are degenerate cases of basic interest. If the flow is in a constant state $r = r_0$, $s = s_0$ in a given region in the r,s-plane, this region will be mapped to only a point in the r, s-plane. There may also be regions in the x,t-plane which map into a line $r = r_0$ (or $s = s_0$) in the r,s-plane. The latter case represents motions referred to as "simple waves".

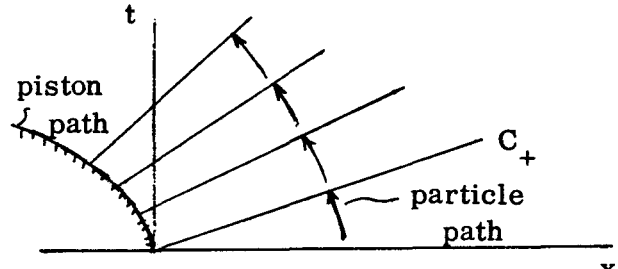
In "simple waves" since the whole region maps to the line $r = r_0$, say, all the s-characteristics (C_--characteristics) become points along $r = r_0$. Back in the x,t-plane, then, along a C_--characteristic $s = s_1$, say, we have

$$r = r_0, \quad s = s_1,$$

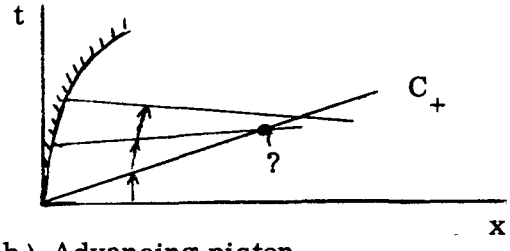
hence ϕ_t and ϕ_x must be constants. The C_--characteristics in the x,t-plane therefore will be straight lines.

Let us now consider the flow which is of constant state in a region of the x,t-plane. This region is mapped to a point (r_0, s_0), say, in the r,s-plane. The boundary between this region of constant state and the adjacent region of variable flow must be a characteristic, say $s = s_0$. Now in the r,s-plane all the s-characteristics must start from (r_0, s_0). The next one $s = s_1$ must be located along the line $r = r_0$, since along the boundary $s = s_0$ the characteristic directions extending into the region of variable flow are still completely specified by $r = r_0$. Thus the adjacent region of variable flow must be mapped into the line segment $\overline{s_0 s_1}$ along $r = r_0$. The conclusion is: The flow adjacent to a region of constant state must be a "simple wave". "Simple wave" solutions consequently are instrumental in constructing solutions containing regions of constant state.

Consider as example the problem of moving a piston in a long tube filled with gas at rest. The bounding characteristic between the region of gas at rest and the region of moving gas is now a C_+. When the piston is retracting, straight C_+-characteristics can be constructed from the prescribed piston path, as in sketch (a), and the flow completely determined. When the piston is advancing, however, the C_+-characteristics so constructed tend to intersect, as in sketch (b). At the intersection we have different values of r and a given s_0, and the values of ϕ_x and ϕ_t can no longer be solved.

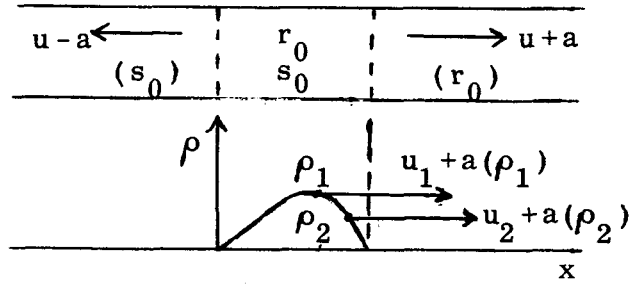


(a) Retracting piston



(b) Advancing piston

The situation is further clarified by considering a slightly different example. Suppose in a long tube of gas at rest a certain portion is disturbed to the state $r_0(x), s_0(x)$ at $t = 0$. Subsequently,

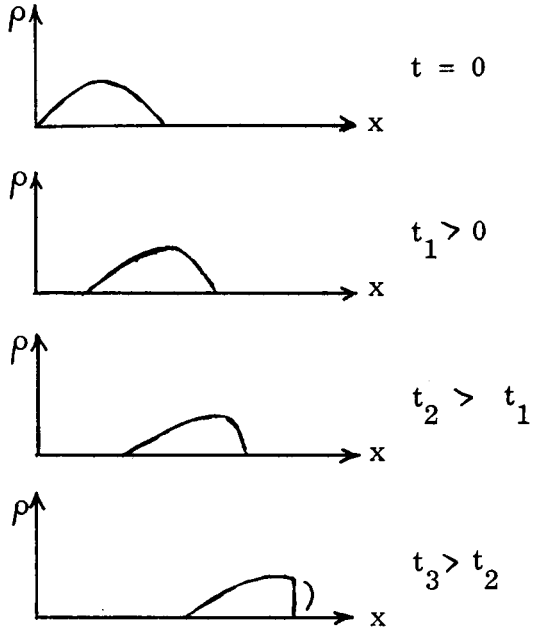


the disturbance r_0 moves to the right at velocity $u + a$, and the disturbance s_0 moves to the left at velocity $u - a$. Let the disturbance r_0 have a density distribution at $t = 0$ as sketched. Since

$$a(\rho) = \sqrt{\gamma p/\rho} \propto \rho^{(\gamma-1)/2}$$

we know that $da/d\rho > 0$. Also, as the boundary between the disturbed and undisturbed regions must be a C_+ - characteristic, the simple wave solution for the forward propagating disturbance satisfies $s = 0$, i.e., $\frac{u}{2} - \frac{a}{\gamma-1} = 0$. Hence $du/da > 0$. Thus if $\rho_1 > \rho_2$ as sketched, we conclude: $u_1 > u_2$ and

$$u_1 + a(\rho_1) > u_2 + a(\rho_2) .$$



The time history of the density disturbance profile will be as shown, with progressive steepening of the "compression side" of the disturbance (increasing density for the fluid element when swept by the disturbance), and progressive flattening of the "expansion side". Eventually it is seen that the simple wave solution must necessarily break down when the profile develops a vertical slope, since any further progress would require the crest to move ahead of the foot, representing a multi - valuedness of the density which is obviously not acceptable. This corresponds to the situation when the characteristics of the same family intersect in the earlier example.

What actually happens in such cases is that discontinuities in the flow variables are developed. The boundary between the disturbed and the undisturbed regions becomes a "shock wave", instead of a characteristic. Without considering the dissipative mechanisms of the viscosity and heat conductivity of the real gas, the shock wave is of zero thickness across which finite changes in u and ρ take place. But the basic conservation laws of mass, momentum and energy for the fluid flow in crossing the shock wave must still be obeyed.

IX. Steady Two-dimensional Homentropic Flows

If we specialize Eq. (VI.10)' to steady two-dimensional flows, the governing equation is

$$(a^2 - \phi_x^2)\phi_{xx} + (a^2 - \phi_y^2)\phi_{yy} - 2\phi_x\phi_y\phi_{xy} = 0 \ . \qquad (IX.1)$$

As in the previous section, to classify this equation, the discriminant will be examined. It reads

$$4\phi_x^2\phi_y^2 - 4(a^2 - \phi_x^2)(a^2 - \phi_y^2) = -4a^2[a^2 - (\phi_x^2 + \phi_y^2)] \ .$$

Thus there are three possibilities:

a) $a^2 - (\phi_x^2 + \phi_y^2) > 0$, i.e., the flow is everywhere subsonic, then the equation is elliptic;

b) $a^2 - (\phi_x^2 + \phi_y^2) < 0$, i.e., the flow is everywhere supersonic, then the equation is hyperbolic;

c) $a^2 - (\phi_x^2 + \phi_y^2)$ changes sign in the flow field, which consists therefore of both subsonic and supersonic regions, then the equation is of the "mixed type".

For subsonic flows, the limiting case of incompressible approximation satisfies $\nabla^2\phi = 0$, which, being linear, can be solved conveniently for most cases. The difficulty of the general case Eq. (IX.1) is primarily in its non-linearity, destroying the possibility of building up a desired solution through superposition. So long as $M^2 < 1$ everywhere, a straightforward procedure is to expand the solution in an ascending power series in, say, the free stream Mach number M_∞, i.e.,

$$\phi = \phi_0 + M_\infty^2\phi_1 + M_\infty^4\phi_2 + \ . \ . \ .$$

where ϕ_0 obviously is the incompressible solution. The successive terms ϕ_1, ϕ_2, etc. satisfy the Poisson equation

$$\nabla^2\phi_n = F_n(\phi_0, \phi_1, \ . \ . \ . \ , \phi_{n-1}) \ .$$

This is known as the Rayleigh-Janzen method. As expected, experience shows that the convergence gets worse when the local Mach number approaches unity somewhere.

More generally, Eq. (IX.1) may be reduced to a linear problem by means of a "hodograph transformation", considering (x, y) as functions of (u, v). The continuity equation, Eq. (VI.3), may be written as

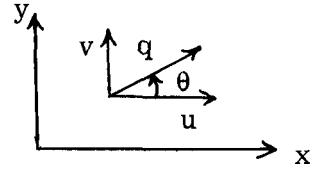
$$\frac{\partial}{\partial x}\rho u + \frac{\partial}{\partial y}\rho v = 0$$

from which a stream function ψ may be defined such that

$$\rho_0 \psi_y = \rho u\,, \quad \rho_0 \psi_x = -\rho v \tag{IX.2}$$

where ρ_0 is a reference constant density. Let now (ϕ, ψ) replace (x, y) as the dependent variables. Let also the the variables (q, θ) in polar coordinates for the velocity, i.e.,

$$\left.\begin{aligned} u &= q\cos\theta \\ v &= q\sin\theta \end{aligned}\right\}$$



replace (u, v) as the independent variables. Then by definition the following complex relation holds,

$$d\phi + i\frac{\rho_0}{\rho}d\psi = u\,dx + v\,dy + i(-v\,dx + u\,dy)$$
$$= q\,e^{-i\theta}dz$$

with $z = x + iy$. Hence

$$z_q = q^{-1}e^{i\theta}\left[\phi_q + i\frac{\rho_0}{\rho}\psi_q\right]$$

$$z_\theta = q^{-1}e^{i\theta}\left[\phi_\theta + i\frac{\rho_0}{\rho}\psi_\theta\right] .$$

Requiring now that $z_{q\theta} = z_{\theta q}$, we find by equating the real and imaginary parts,

$$\left.\begin{aligned} \phi_q &= q\frac{d}{dq}\left[\frac{\rho_0}{\rho q}\right]\psi_\theta \\ \phi_\theta &= \frac{\rho_0 q}{\rho}\psi_q \end{aligned}\right\} \tag{IX.3}$$

ρ being here regarded as a function of q. It is now possible to derive a <u>linear</u> equation in either ϕ or ψ by elimination. For instance, in terms of ψ,

$$\frac{\partial}{\partial\theta}\left[q\frac{d}{dq}\left(\frac{\rho_0}{\rho q}\right)\psi_\theta\right] - \frac{\partial}{\partial q}\left[\frac{\rho_0 q}{\rho}\psi_q\right] = 0. \tag{IX.4}$$

This equation was first derived by Chaplygin in 1904 in his investigation on gas jets. The disadvantage here is that the boundary conditions involving a given body become very involved. One usually has to take a solution and then find out the exact body shape for which it is the solution.

The relation $\rho(q)$ implied above of course is given by the Bernoulli equation.

Now in the incompressible case, Eq. (IX.3) reduces to

$$\phi_q = -\frac{1}{q}\psi_\theta, \quad \phi_\theta = q\psi_q .$$

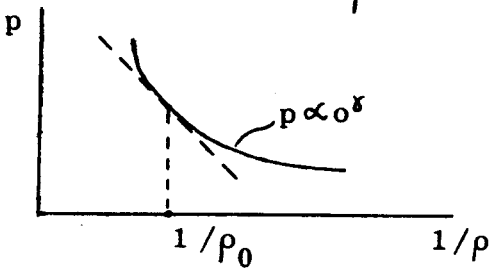
Chaplygin observed that the general case will assume a similar form if

$$q\frac{d}{dq}\left(\frac{\rho_0}{\rho q}\right) = -\frac{\rho}{\rho_0 q}$$

which may be integrated into $q^2 \propto 1 - \left(\frac{\rho_0}{\rho}\right)^2$, expressing the required $\rho(q)$. Indeed the Bernoulli equation yields such a form for the hypothetical gas with $\gamma = -1$. Thus by approximating the time isentropic relation $p \propto \rho^\gamma$ by an expression

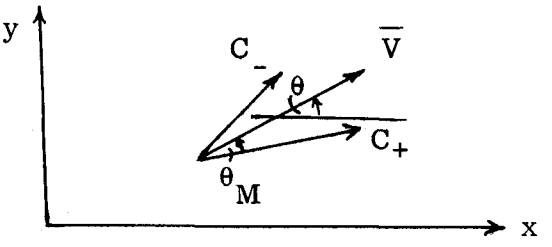
$$p = a + b\rho^{-\gamma}$$

it becomes possible to relate incompressible and compressible solutions in the hodographic variables. The well-known Kármán-Tsien approximation for subsonic flows amounts to a tangent approximation of the isentropic curve p vs. $\frac{1}{\rho}$, near the reference density ρ_0, chosen as the density at the "stagnation point" where $q = 0$.



It is noted that Eq. (IX.3) is no less general that Eq. (IX.1) and therefore not restricted to subsonic flows. As a matter of fact, certain simple solutions give examples of continuous flows involving both subsonic and supersonic regions. In interchanging the roles of (x, y) and (q, θ), however, a one-to-one correspondence is implied. When $\partial(q, \theta)/\partial(x, y) = 0$ a finite region in the x, y-plane is mapped to a line or a point in the q, θ-plane. We get for the former a solution, if $M > 1$ everywhere, known as the "Prandtl-Meyer flow", corresponding to the "simple wave" of the previous section, while the latter clearly represents a uniform flow. When $\partial(x,y)/\partial(q,\theta) = 0$, a finite region in the hodograph plane is mapped generally to a line in the physical plane, requiring therefore multi-valuedness of the flow along this line -- again a physically unacceptable situation. Such lines are known as "limit lines", occurring only when $M > 1$ locally and indicating the breakdown of the assumed continuous irrotational homentropic flow.

In the region of supersonic flow, since Eq. (IX.1) becomes hyperbolic, the method of characteristics again may be used. The characteristic direction, corresponding to Eq. (VIII.3) are found to be



$$\frac{dy}{dx} = \frac{-\frac{uv}{a^2} \mp \sqrt{M^2 - 1}}{1 - \frac{u^2}{a^2}} \qquad (IX.5)$$

the "$\mp$" sign corresponding to the $C_{\pm}$-directions, resp. in the sketch. It can be verified that the $C_{\pm}$-directions make an angle θ_M with the local velocity vector, θ_M being the "Mach angle",

$$\theta_M = \sin^{-1} \frac{1}{M} .$$

The characteristic conditions, corresponding to Eq. (VIII.4), may be conveniently expressed in p and θ, θ denoting the local velocity direction, as

$$\frac{\cot \theta_M}{\rho} dp \mp q^2 d\theta = 0 \quad \text{along} \quad C_{\pm} . \qquad (IX.6)$$

The "simple wave" solution when a finite region if the x, y-plane is mapped to a single characteristic C_+, say, in the p, θ-plane follows directly

$$dp = \rho q^2 d\theta / \sqrt{M^2 - 1} ,$$

hence $dp/d\theta > 0$ in such flows. In conjunction with the Bernoulli equation the above may be integrated. We only note that since

$$\frac{dp}{\rho} + q\,dq = 0$$

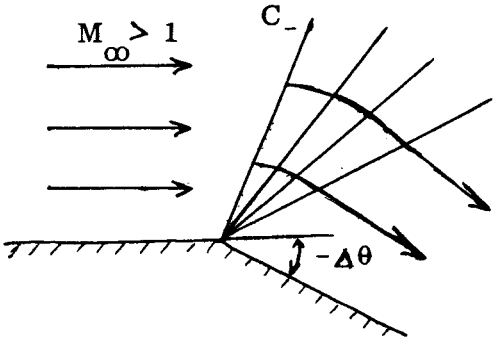
the "simple wave" equation may also be written as

$$\frac{dq}{q} = -\frac{d\theta}{\sqrt{M^2 - 1}}$$

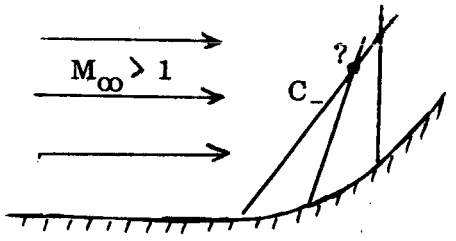
hence $dq/d\theta < 0$ in such flows. Thus speed increases as pressure (or density) decreases, and vice versa.

The same argument in the previous section may be followed to prove that a region of uniform supersonic flow can be extended continuously into a region of

variable flow only through the simple wave solution. As an example, the supersonic flow turning around a corner is obtained by drawing successive C_- - lines from the corner until the velocity leaving the last C_- - line has turned through the full angle $\Delta\theta$. Since θ is continuously decreasing in the stream direction, pressure drops and speed goes up as a result. The transition region is known as the Prandtl - Meyer expansion fan.

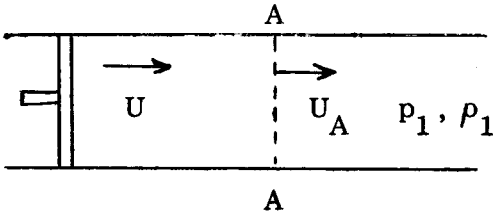


If the flow is along a concave wall, $\Delta\theta > 0$ and pressure tends to rise in the streamwise direction. Here again the C_- - characteristics will intersect and a continuous solution becomes impossible. Thus we must expect shock waves to appear in two - dimensional steady flows when a supersonic stream is subjected to a compressive disturbance (increasing pressure in the streamwise direction).



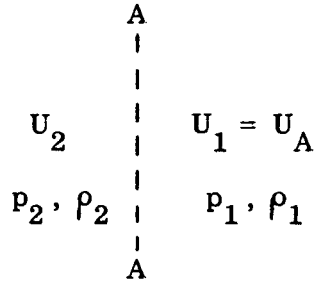
X. Shock Conditions and Flows with Shocks

Consider now the one - dimensional problem of a piston moving uniformly at velocity U into a long tube containing gas at rest with pressure p_1 and density ρ_1. We postulate a shock wave separating the disturbed and undisturbed regions, advancing at an unknown velocity U_A. It seems clear that U_A must be defined by only U, p_1, and ρ_1, hence a constant. By dimensional reasoning,



$$U_A = U\,F\left(\frac{U}{\sqrt{p_1/\rho_1}}\right).$$

Now it is possible to let an observer ride on the shock AA, reducing the flow near the shock to a steady one. For a small area on AA, the conservation laws of mass, momentum and energy then give



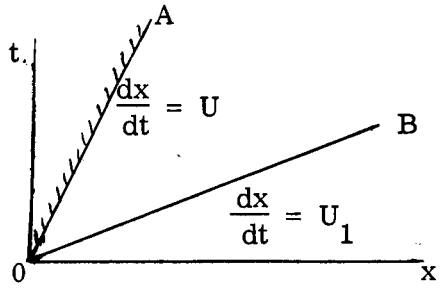
$$\left.\begin{aligned} &\rho_1 U_1 = \rho_2 U_2 \\ &p_1 + \rho_1 U_1^2 = p_2 + \rho_2 U_2^2 \\ &\rho_1 U_1 \left(h_1 + \frac{U_1^2}{2}\right) = \rho_2 U_2 \left(h_2 + \frac{U_2^2}{2}\right) \end{aligned}\right\} \qquad (X.1)$$

h being the enthalpy as used in Eq. (IV.14)". Together with the equation of state, $p = \rho RT$, all the variables with suffix "2" can be solved in terms of those with suffix "1". The result is known as the "Rankine-Hugoniot" relations. It turns out that for given

$$M_1 \equiv U_1 / \sqrt{\gamma RT_1} > 1 \ , \quad p_2/p_1, \ \rho_2/\rho_1 \ \text{ and } \ T_2/T_1$$

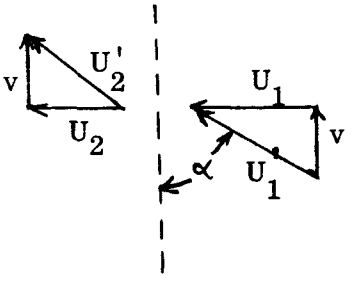
• are all greater than unity, while $M_2 \equiv U_2 / \sqrt{\gamma RT_2} < 1$. Furthermore, the entropy S_2 is found to be higher than S_1. Hence, although the same Eq. (X.1) holds if both U_1 and U_2 are reversed in direction, the second law of thermodynamics is obeyed only when the motion is in the direction sketched. Thus the shock wave propagates into the calm region at a supersonic speed (with respect to the sound speed ahead of it).

Turning back to the piston problem, we see that a solution is possible by taking the disturbed region to be in a uniform state, determined by requiring $U_2 = U - U_1$ after the shock of the Mach number M_1. That there is no other solution can also be shown in the following manner: We need a solution in the wedge shaped region in the x, t-plane as shown, taking on the velocity $u = U$ along the line OA and the velocity $U_2 - U_1$ and density ρ_2 from the shock relation along the line OB. Now the lack of a length scale suggests that the solution must depend on a single variable $\xi = x/t$. It is then easily verified that a solution of the form $\rho = \rho(\xi)$, $u = u(\xi)$ cannot be made to satisfy the boundary condition except as constants.

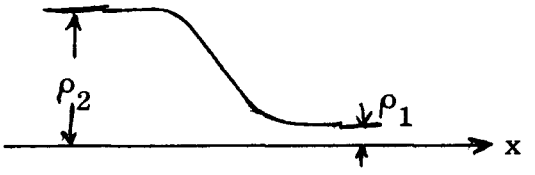


It further is clear that a uniform translation of the entire flow field in any direction should not affect the conservation laws. In particular, by imposing a uniform velocity U parallel to the wave front AA, the oblique shock making an angle κ with the

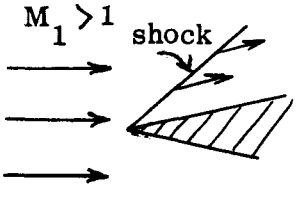
oncoming velocity U_1' is obtained. The normal velocity component is therefore seen to be the effective one in causing the shock wave. In this way the oblique shock relations follow immediately. As the conservation laws are applied to a small area on the wave front in deriving Eq. (X.1), the shock relations are actually local in nature and remain valid locally on any curved shock surface.

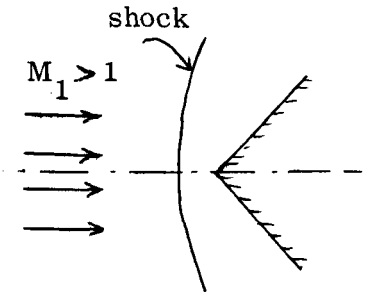


The real gas of course has viscosity and heat conductivity as dissipative mechanisms, which resist the discontinuity occurring in a shock wave of zero thickness. The net effect is to smear out the shock wave, so that the upstream and downstream densities ρ_1 and ρ_2, say, are approached only asymptotically. However, most of the change occurs in a very small thickness of the order of the mean free path of the gas molecules. Unless the upstream or downstream part of the flow varies significantly in such a small thickness, the shock structure plays a negligible role in fluid dynamics.



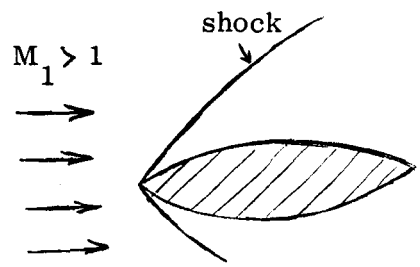
We now mention some examples of steady flows with shock waves. Consider a two-dimensional wedge placed in a uniform supersonic stream. In previous sections, we concluded that the compressive disturbances due to the turning of the streamlines to parallel the wedge surface causes the presence of shock waves. By inserting a straight oblique shock attached to the vertex, it is generally possible to have a uniform and supersonic flow parallel to the wedge after the shock. For a given $M_1 > 1$, however, the wedge angle may be too large for any attached straight shock to turn the streamline sufficiently. Then it becomes necessary to postulate a "detached shock" in front of the body, starting necessarily as a normal shock at the line of symmetry. The flow behind the normal shock is of course subsonic, but as the shock bends gradually toward the body surface away from the line



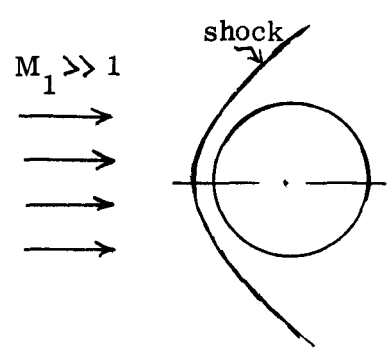


of symmetry, the flow behind the shock eventually becomes supersonic. The flow problem involving detached shocks is thus of the mixed type and difficult to treat except numerically.

As soon as curved shocks appear in a uniform stream, it should be noted that the flow behind it is strictly speaking always rotational. The entropy change depends on the obliqueness of the shock, and the different values of entropy along different streamlines give rise to vorticity (Eq. (IV.7), "Crocco Theorem"). For instance, to calculate the flow around a two-dimensional curved body with pointed nose in a supersonic stream, the solution should be started with an attached shock at the nose and continued by the method of characteristics, complicated by the unknown shock inclination at successive steps as well as the resulting rotational nature of the flow. However, the entropy change across a shock turns out to be third order in the "shock strength" parameter, which may be taken as $(p_2 - p_1)/p_1$. If the shock is not strong, the assumption of isentropic flow is not too far wrong. Thus a practical approximate method for the curved body problem, known as the "shock expansion method", is to regard the streamline immediately adjacent to the body as following a Prandtl-Meyer expansion after the leading edge shock. Its use of course cannot be extended to hypersonic flows where the shock will always be quite strong.



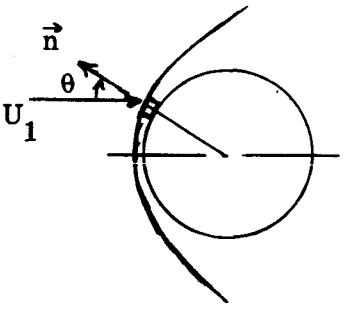
In hypersonic flow the blunt body is of practical interest. The very difficult problem of the mixed type flow behind a detached shock is an inherent feature. The limiting case of $M_1 \to \infty$, however, permits at least a much simplified first approximation. Based upon the observation that, after a normal shock the Rankine-Hugoniot conditions give the density ratio as



$$\frac{\rho_2}{\rho_1} = \frac{U_1}{U_2} = \frac{(\gamma+1)\,M_1^2}{2 + (\gamma-1)\,M_1^2} \rightarrow \frac{(\gamma+1)}{(\gamma-1)} \quad \text{as} \quad M_1^2 \rightarrow \infty,$$

hence if $\gamma = 1$, $\rho_2/\rho_1 \rightarrow \infty$. For air at normal temperatures $\gamma = 1.4$; at very high temperatures with dissociation γ becomes even closer to unity. Thus as a rough approximation we might examine the flow with $\gamma = 1$. Now as $\rho_2/\rho_1 \rightarrow \infty$, the shock will locally simply wrap around the body, since the continuity equation is only satisfied by having no thickness between the shock and the body. Neglecting the actual thin "shock layer" thickness, the pressure on the body can be determined directly from simple momentum consideration:

$$p - p_1 \cong \rho_1 U_1^2 \cos^2\theta.$$



This result is identical with what would have been predicted according to Newton's corpuscular theory, that the oncoming gas consists of particles moving at the same speed U_1. Hence this type of approximation is referred to as "Newtonian". For refinement the effects of the error of $\gamma = 1$ has been accounted for, for instance, by expanding the solution in terms of the small parameter $(\gamma - 1)/(\gamma + 1)$.

XI. Viscous Flows and the Low Reynolds Number Approximations

We have so far considered a great deal of the fluid motion under the inviscid approximations, on the basis that for small viscosity and high speeds the viscous effects will be confined to a thin boundary layer immediately adjacent to the body and to a wake behind the body. Such theories obviously can be of no value in connection with the question of skin friction and heat transfer which depend on the details of the motion within the boundary layer. Furthermore, the precise boundaries of the wake must be known in order to construct the essentially inviscid solution surrounding the body and the wake.

It was first systematically observed by Reynolds in pipe flows that the viscous fluid motion can assume different forms dependent essentially on the dimensionless Reynolds number $Re \equiv VL/\nu$, where V is the characteristic velocity (e.g., mean flow velocity through the pipe), L a characteristic length

(e.g., the pipe diameter) and ν the kinematic viscosity of the fluid. On the one extreme, at sufficiently low Re the fluid may move steadily in layers parallel to the pipe axis; on the other, at sufficiently high Re the motion may become time-dependent, irregular and random, but with well defined time averages. The former is referred to as "laminar" motion while the latter is referred to as "fully turbulent" motion. Naturally there is also a "transition" region in which the laminar motion develops into the fully turbulent one. Similar types of motion prevail also in boundary layers. The details of the flow clearly cannot be investigated without first knowing whether the motion is laminar or turbulent. The question of the transition from laminar to turbulent motion is thus of prime importance.

Generally speaking, when the flow is ostensibly governed by physical parameters which are invariant with time, the laminar motion corresponds to the solution of the equations of motion under the assumption of steady flow. If deviations from the steady flow can occur without changing any of the governing physical parameters, the question must be one of stability. As usual the stability problem may be formulated by studying the behavior of perturbations. Unfortunately, mathematically the stability theory is rather difficult even for very simple laminar flows under infinitesimal disturbances. The results are further valid only for the initial breakdown of the laminar flow. Nevertheless the stability theory does provide qualitative correlations between transition and the various physical parameters. The actual beginning of the fully turbulent region however is yet beyond the capability of theoretical prediction. The situation is complicated in addition by the fact that, for bodies in flight, irregularities of the body surfaces and in the free stream all have profound influence on transition.

The analysis of fully turbulent flow is even more difficult. By putting the instantaneous flow variable as the sum of the "mean" part plus a fluctuation, equations for the mean motion may be derived from the general equations of motion, but contributions due to the non-linear interaction of various fluctuations inevitably show up as additional unknowns. For instance, in the mean momentum equation, the momentum transfer due to the fluctuating velocity components through the fixed control surfaces of a fluid element leads to the turbulent or "Reynolds" stresses.

In simplified analyses, ad hoc assumptions are made by expressing the Reynolds stresses in terms of other mean flow variables, and a formal solution may then be carried out involving adjustable parameters, which are finally chosen in some way to agree with experimental findings. Such theories are of course semi-empirical in nature, but often unavoidable for practical purposes.

We restrict ourselves in the following to only some of the laminar flow problems. It may be noted that most of the peculiar nature of viscous fluid motion owes to the relative roles of the viscous term and the non-linear convective terms such as $\vec{V} \cdot \nabla \vec{V}$ in the equations of motion. Thus not much generality is lost when the complications of compressibility are omitted for brevity. For an incompressible viscous fluid, Eqs. (III.2) and (III.3), together with Eqs. (III.9) and (III.11), lead to the following

$$\left.\begin{aligned} &\nabla \cdot \vec{V} = 0 \\ &\rho \left(\frac{\partial \vec{V}}{\partial t} + \vec{V} \cdot \nabla \vec{V} \right) = -\nabla p + \mu \nabla^2 \vec{V} \end{aligned}\right\} \qquad \text{(XI.1)}$$

known as the "Navier-Stokes equations", in which the viscosity coefficient may be regarded as constant if the temperature range is small. Our purpose is to examine some of its solutions for flows over bodies. The boundary condition on the body is that of "no slip" as discussed in §IV. After the velocity and pressure fields are determined, the temperature field may then be solved separately from the energy equation Eq. (III.14), under the "no jump" boundary condition.

Since Eq. (XI.1) is non-linear, an attempt to simplify is naturally that of linearization for small perturbations. Considering therefore an object moving at very low speed in a fluid at rest, we might neglect the quadratic "convective" terms $\vec{V} \cdot \nabla \vec{V}$ from Eq. (XI.1). It follows immediately that

$$\left.\begin{aligned} &\nabla^2 p = 0 \\ &\nabla^2 \nabla^2 \nabla = 0 \, . \end{aligned}\right\} \qquad \text{(XI.2)}$$

Since the highest order derivatives are not disturbed, it appears that all the boundary conditions for the original equations can be accomodated. This was first used by Stokes to calculate the drag on a sphere moving steadily at a low speed V_∞ in an unbound fluid at rest, and is known as "Stokes' approximation". By using the

sphere radius "a" as a characteristic length and the speed V_∞ as the characteristic velocity, an order of magnitude estimate gives

$$\frac{\rho \vec{V} \cdot \nabla \vec{V}}{\mu \nabla^2 \vec{V}} \sim \frac{V_\infty a}{\nu} = Re_a .$$

Thus Stokes' approximation corresponds to the limiting case of $Re_a \ll 1$, i.e., very low Reynolds numbers based upon the sphere radius "a".

The explicit solution of Stokes' sphere problem, however, leads to asymptotically for large r,

$$\vec{V} \sim \vec{V}_\infty \left[1 + O\left(\frac{a}{r}\right) \right]$$

where r is the radial distance measured from the center of the sphere, the coordinate axis having been fixed on the sphere. As $r/a \to \infty$, we find in fact

$$\frac{\rho \vec{V} \cdot \nabla \vec{V}}{\mu \nabla^2 \vec{V}} \sim \frac{V_\infty r}{\nu} \to \infty ,$$

showing that Eq. (XI. 2) cannot help but fail as an approximation of Eq. (XI. 1) at far enough distances, regardless of the smallness of the Reynolds number. In other words, the convective terms eventually take the upper hand as compared with the viscous terms. The seemingly innocent Stokes' approximation is not uniformly valid. In fact, for two-dimensional problems, it is easy to see that Eq. (XI. 2) must lead to asymptotically for large r,

$$\vec{V} \sim \vec{V}_\infty \left[1 + O\left(\log \frac{a}{r}\right) \right].$$

In an unbound fluid the condition of uniform stream at large distances cannot be satisfied at all.

The difficulty of Stokes' approximation is perhaps best understood by observing that there are actually <u>two</u> characteristic lengths in viscous flow. Besides the geometrical length "a" we also have a viscous length ν/V_∞. In the near field close to the body, the dimensionless distance of interest is indeed r/a, but in the far field away from the body the flow must be expressible in terms of the dimensionless distance $\nu r/V_\infty$ regardless of the body. In general, therefore, two separate approximations are called for, to be matched somehow in an overlapping region where both might be acceptable. In the terms of Lagerstrom, Kaplan and Van Dyke, it is an example requiring the matching of an "inner" and an "outer"

expansion.*

The criticism of Stokes' approximation regarding its behavior at large distances from the body was first made by Oseen. As a remedy, Oseen's proposal was to recognize the far field as a small perturbation of the steady uniform stream, hence

$$\vec{V} \cdot \nabla \vec{V} \cong \vec{V}_\infty \cdot \nabla \vec{V}' \qquad (XI.3)$$

where $\vec{V}' = \vec{V} - \vec{V}_\infty$, the perturbation velocity. Since the rest of the terms in Eq. (XI.1) are linear in $\vec{V}$, they may all be written without change in terms of $\vec{V}'$. Retaining Eq. (XI.3) as a first approximation of the convective terms everywhere, we get the "Oseen equation" for steady flows -

$$\left.\begin{aligned} \nabla \cdot \vec{V}' &= 0 \\ \rho \vec{V}_\infty \cdot \nabla \vec{V}' &= -\nabla p + \mu \nabla^2 \vec{V}' \end{aligned}\right\} \qquad (XI.4)$$

again with the same boundary condition as before but expressed in $\vec{V}'$. As an approximation for the far field, evidently the primary effects of the convective terms are represented correctly. As the body is approached, the flow will be characterized by the geometrical length "a" and the convective terms still are much smaller than the viscous terms for $V_\infty a/\nu \ll 1$. It is however of uncertain validity in the region where the two types of terms are comparable. For the sphere problem, the drag coefficient from the Oseen approximation is found to be

$$C_D = \frac{6\pi}{Re_a}\left[1 + \frac{3}{8} Re_a + O(Re_a^2)\right]$$

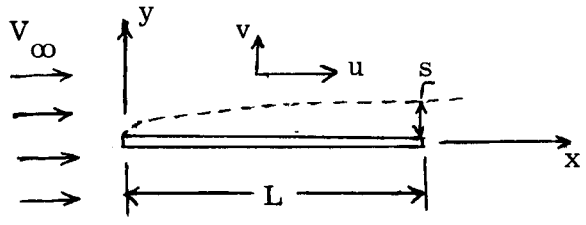
where the first term agrees with Stokes' result. (Terms up to $O(Re_a^6)$ have been computed by Goldstein.) A recent more careful analysis shows that the $O(Re_a^2)$ term in the bracket actually should be $\frac{9}{40} Re_a^2 \log Re_a$ (Proudman and Pearson, J. Fluid Mech. 2, 237-262, 1957).

XII. Theory of the Boundary Layer

In the other extreme of large Re, we need to describe the motion in the thin "boundary layer" immediately adjacent to the body. In §V it is seen that the boundary

* For the technique of "inner" and "outer" expansions, see e.g. M. van Dyke: "Perturbation Methods in Fluid Mechanics", Lecture Notes, Stanford University.

layer thickness δ is $O(1/\sqrt{Re_L})$. For a point fixed in space, no matter how close to the body surface, as $Re \to \infty$ it will lie outside of the boundary layer. This corresponds to dropping formally the viscous terms in the Navier-Stokes equation, and yields the inviscid approximation. In order to keep the point in question within the boundary layer, we must therefore maintain y/δ finite, where y denotes the distance from the body surface, even as $\delta \to 0$ in the limit. To be more specific, consider for simplicity the steady two-dimension motion of a flat plate moving parallel to itself in an unformed incompressible fluid. We have here again two length parameters: the geometric length L of the plate, and the viscous boundary layer thickness δ, $\delta \sim L/\sqrt{Re_L}$. In the limit $Re_L \to \infty$ or $\delta \to 0$, the inviscid solution is simply the undisturbed uniform flow. The u-component velocity in the boundary layer parallel to the plate is generally characterized by V_∞. The order of magnitude of the v-component velocity may be inferred from the continuity equation,



$$v \sim O\left(\int_0^\delta \frac{\partial u}{\partial x}\, dy\right) \sim O\left(\delta \frac{\partial u}{\partial x}\right) .$$

Since all changes in the y-direction must be accomplished within the thickness δ, there follows also

$$\frac{\partial}{\partial y} \sim O\left(\frac{L}{\delta}\right) \frac{\partial}{\partial x} \sim O\left(\sqrt{Re_L}\right) \frac{\partial}{\partial x} .$$

Thus by introducing dimensionless variables of comparable magnitudes

$$u^* = u/V_\infty \quad , \quad v^* = (v/V_\infty)\sqrt{Re_L} \quad , \quad p^* = p/\rho V_\infty^2$$

$$x^* = x/L \quad , \quad y^* = (y/L)\sqrt{Re_L} \, ,$$

Eq. (XI.1) becomes

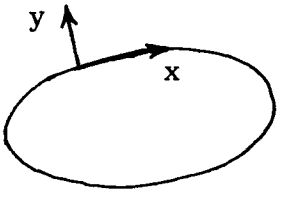
$$\left.\begin{aligned} &\frac{\partial u^*}{\partial x^*} + \frac{\partial v^*}{\partial y^*} = 0 \\ &u^* \frac{\partial u^*}{\partial x^*} + v^* \frac{\partial u^*}{\partial y^*} = -\frac{\partial p^*}{\partial x^*} + \frac{\partial^2 u^*}{\partial y^{*2}} + O\left(\frac{1}{Re_L}\right) \\ &O\left(\frac{1}{Re_L}\right) = -\frac{\partial p^*}{\partial y^*} . \end{aligned}\right\} \qquad (XII.1)$$

We now let $Re_L \to \infty$ and omit the terms $O(1/Re_L)$. The result is the boundary layer equation of Prandtl. The most important feature is the replacement of $\nabla^2 u$ by $\partial^2 u/\partial y^2$ in the x-momentum equation, retaining at least one of the highest order derivatives in the full equation. The solution is also greatly simplified by the consequence of the approximate y-momentum equation that leads to

$$p^* = p^*(x^*),$$

i.e., constant pressure across the boundary layer at each streamwise station. In the flat plate case under consideration the pressure must agree with that in the free stream, hence a constant. The derivation, however, obviously is not valid wherever u and v are of the same order of magnitude, such as the stagnation point regions at the leading and trailing edges of the flat plate.

The same order of magnitude arguments can be applied to bodies of arbitrary but smooth shape. By interpreting the x-coordinate as running along the body surface and y normal to it, the same boundary layer equations result except that the omitted terms include those of $O(\kappa\delta)$, where κ is the characteristic curvature of the body shape. The pressure remains unchanged in the y-direction, the correction due to centrifugal force being $O(\kappa\delta)$. To match the boundary layer solution with the inviscid solution which prevails beyond the boundary layer, it is noted that since $\delta \to 0$ in the limit, the conditions at the "edge of the boundary layer" must agree with the inviscid solution evaluated at the body surface. This consideration leads to the boundary conditions that, as $y^* \to \infty$



$$u \to u_i(x,0), \quad p \to p_i(x,0)$$

$u_i(x,y)$ and $p_i(x,y)$ being the inviscid solutions. There is on the other hand no condition on v^* as $y^* \to \infty$; so long as it is finite, the discrepancy between $v = \lim_{Re_L \to \infty} v^*/\sqrt{Re_L}$ and $v_i(x,0) = 0$ is of no consequence at this level of approximation. At the body surface $y = 0$, the "no slip" condition of viscous fluids must be satisfied by setting $u = 0$, $v = 0$ as usual.

It should be remarked that the boundary layer equation would assume different

forms depending on the choice of the coordinate system, hence also the flow field which follows as the solution. In his study of the two-dimensional steady incompressible boundary layers, Kaplan introduced the notion of an "optimal" system of coordinates that render the boundary layer solution to agree completely, as $y^* \to \infty$, with the inviscid solution evaluated at the surface $y \to 0$, in both u- and v-components to $O(1/\sqrt{Re_L})$. But the boundary layer solutions in the optimal and any other non-optimal system of coordinates are shown to be able to transform into each other. Furthermore, he proved that the skin friction at the body surface is independent of the coordinate system. The choice of the coordinate system is therefore not too crucial for ordinary purposes.

Though much simplified from the full Navier-Stokes equation, the non-linear boundary layer equation still defies general treatment. To reduce Eq. (XII.1) to a single dependent variable, the stream function ψ may be introduced by defining

$$u^* = \partial\psi/\partial y^* \quad , \quad v^* = -\partial\psi/\partial x^*$$

guaranteeing thereby the satisfaction of the continuity equation. Now we apply the "von Mises transformation" to the second equation of Eq. (XI.1) by choosing (x, ψ) as the independent variables instead of (x, y) and obtain

$$\left.\begin{aligned} u^* \frac{\partial u^*}{\partial x^*} &= -\frac{dp^*}{dx^*} + u^* \frac{\partial}{\partial \psi}\left(u^* \frac{\partial u^*}{\partial \psi}\right) \\ p^* &= p^*(x) \text{, given} \quad . \end{aligned}\right\} \qquad \text{(XII.2)}$$

This equation is clearly parabolic in nature. Indeed, if the dimensionless stagnation pressure P^* replaces u^* as the dependent variable

$$P^* \equiv p^* + \frac{u^{*2}}{2}$$

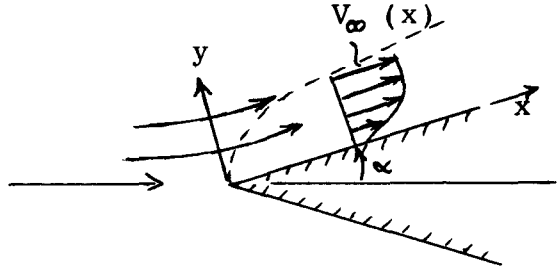
Eq. (XII.2) may be put into the form

$$\frac{\partial}{\partial x^*} P^* = [P^* - p^*]^{y_2} \frac{\partial^2 P^*}{\partial \psi^2}$$

which becomes the diffusion equation but with a complicated diffusivity. Thus the solution $P^*(x, \psi)$ can be found in the region $x^* \geq 0$, $\psi \geq 0$ if we specify an initial profile $P^*(0, \psi)$ along $x^* = 0$ and also the condition $P^*(x, 0)$ along $\psi = 0$. Since $p^*(x^*)$ is a given function, the conditions on P^* of course are equivalent to the statement that from a given initial velocity profile $u(0, y)$ the

solution can be continued uniquely to $x > 0$ for given $p(x)$. As $\psi \to \infty$, $\partial^2 P^*/\partial\psi^2 \to 0$, hence $\partial P^*/\partial x \to 0$ and the solution merges with the inviscid potential flow $P^* = \text{const}$. Note also that no disturbances are propagated upstream.

It is of interest to be able to stipulate an initial profile for the boundary layer over an arbitrary body. For all blunt-nosed bodies there always exists a front stagnation point O. Locally the flow is equivalent to that against an infinite wall normal to the stream. This is recognized as a special case of the symmetrical flow against a wedge of half angle α. (In fact, the flat plate also is a special case, with $\alpha = 0$.) In all such cases it turns out that the inviscid potential flow is of the type



$$V_\infty \propto x^m,$$

the exponent m depending on the half angle α, $0 \le m \le 1$ for $0 \le \alpha \le \pi/2$. Although we do not have an initial profile $u(0,y)$, a class of "similar solutions" can be found in the form

$$\frac{u}{V_\infty(x)} = \frac{dF}{d\eta}$$

where $F = F(\eta)$ and η is defined by

$$\eta = y/[2\nu x/(1+m)\,V_\infty]^{1/2}.$$

The function $F(\eta)$ corresponds to the stream function, satisfying an ordinary differential equation:

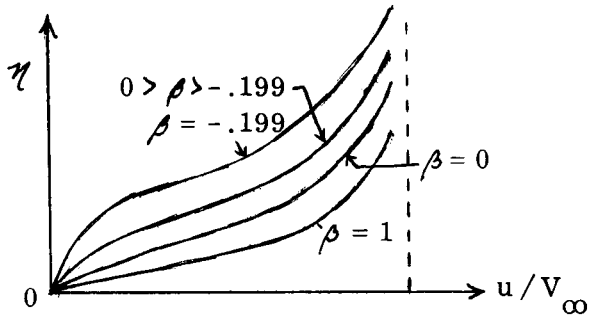
$$\left.\begin{aligned} &F''' + FF'' + \beta(1 - F'^2) = 0 \\ &\beta = 2m/(1+m) \end{aligned}\right\} \qquad \text{(XII.3)}$$

with the required boundary conditions:

$$\left.\begin{aligned} &y = 0,\ u = v = 0 \quad \text{or} \quad F(0) = F'(0) = 0 \\ &y \to \infty,\ u \to V_\infty(x) \quad \text{or} \quad F'(\infty) \to 1. \end{aligned}\right\} \qquad \text{(XII.4)}$$

The case of $m = 0$ gives the Blasius solution for the flat plate. The case of $m = 1$ gives the stagnation point solution, which is in fact the exact solution of the Navier-Stokes equation for the same problem. The existence and uniqueness of the solution

of Eq. (XII. 3) under boundary condition Eq. (XII. 4) are established mathematically for $\beta \geq 0$ by Weyl. Numerical solutions for various values of β (or m) were calculated by Falkner and Skan, and refined by Hartree. For $\beta < 0$, it is interesting to note that the condition $F'(\infty) \to 1$ fails to determine a unique solution, and Hartree had to stipulate the additional requirement that $F'(\eta)$ should approach unity in the most rapidly possible way. Even so,he had to stop at $\beta = -.199$, beyond which the velocity develops an overshoot within the boundary layer which seems physically hard to accept. At $\beta = -.199$ the profile has the feature that $\partial u/\partial y\big|_{y=0} = 0$.



The similar solutions are useful not only to start numerical calculations near the stagnation point, but have often been used as the basis for constructing a first approximation for flows involving rather arbitrary pressure distributions as might occur in practical problems. Let an arbitrary $V_\infty(x)$ be given, then $-dp/dx = V_\infty(dV_\infty/dx)$. Now, proceeding as for similar solutions, let $u/V_\infty = \partial F/\partial\eta$, but $F = F(\xi, \eta)$ with

$$\xi \equiv \frac{1}{\nu}\int_0^x V_\infty\, dx, \quad \eta \equiv y/\left[\frac{\nu}{V_\infty}\sqrt{2\xi}\right].$$

The boundary layer equation for F in (ξ, η) is found to be

$$\left.\begin{aligned} &F_{\eta\eta\eta} + F F_{\eta\eta} + \beta(\xi)(1 - F_\eta^2) = 2\xi(F_\eta F_{\xi\eta} - F_\xi F_{\eta\eta}) \\ &\beta = \frac{2\xi}{V_\infty^2}\frac{dV_\infty}{dx}. \end{aligned}\right\} \quad \text{(XII. 5)}$$

The boundary conditions are still Eq. (XII. 4)

$$\eta = 0,\ F = F_\eta = 0; \quad \eta \to \infty,\ F_\eta \to 1.$$

The approximation next is to neglect the right-hand side and solve F as an ordinary differential equation with $\beta(\xi)$ as a parameter. In other words, the similar solution corresponding to the local $\beta(\xi)$ is used as an approximation, the past history being partially accounted for by the ξ-transformation. This is known as the "locally similar" approximation. For improvement Görtler took it as the first term of a series expansion in the solution of Eq. (XII. 5). Nickel (K. Nickel; Ing. Arch. 31,

85-100, 1962) verified that the local similar solution always provides a lower bound of the true solution $u(x,y)$ so long as $d\beta/d\xi \leq 0$.

Let us adopt the locally similar approximation to get a qualitative picture of the flow within a boundary layer under pressure gradient. If the pressure gradient is always "favorable" ($V'_\infty \geq 0$), $\beta \geq 0$, we expect rather normal velocity distributions somewhat like that on a flat plate. But if $V'_\infty < 0$, it becomes possible for β to reach the critical value of $-.199$ and even exceed it. There can then be no description of the boundary layer beyond that point. The question arises: what happens then?

It is usually taken that what happens then is the observed phenomenon of "separation", i.e., the streamline begins to detach from the surface. Beyond the separation point, close to the surface the flow direction will be reversed, and the boundary layer approximation ceases to hold. That separation indeed could happen at $\beta = -.199$ is made plausible by noting that here $\partial u/\partial y\big|_{y=0} = 0$. Since the streamline direction at the wall is given by

$$\frac{dy}{dx} = \frac{v}{u}\bigg|_{y=0} = \frac{\partial v/\partial y}{\partial u/\partial y}\bigg|_{y=0} = -\frac{\partial u/\partial x}{\partial u/\partial y}\bigg|_{y=0}$$

hence $dy/dx = 0$ if $\partial u/\partial y\big|_{y=0} \neq 0$; but the slope is of the form $0/0$, therefore indefinite if $\partial u/\partial y\big|_{y=0} = 0$. According to this criterion, an adverse pressure gradient is necessary for separation, in perfect agreement with experience.

We briefly turn to the energy equation again for incompressible fluids, to introduce the basic concepts in heat transfer. The equation reads, after taking the boundary layer approximations,

$$\rho u \frac{\partial h}{\partial x} + \rho v \frac{\partial h}{\partial y} = \frac{\partial}{\partial y}\left(\frac{k}{C\rho}\frac{\partial h}{\partial y}\right) = \frac{\mu}{Pr}\frac{\partial^2 h}{\partial y^2} \qquad \text{(XII.6)}$$

where the Prandtl number $Pr \equiv \mu/C\rho k \cong .76$ for air in ordinary temperature range. The equation clearly is of the same structure as the x-momentum equation in Eq. (XII.1). When (u,v) are described by a similar solution, h will

resemble u in behavior and depend on the same variable. The equation is linear for given (u,v), and can be more easily solved in general. Let us mention only the simplest case of flat plate, where the x-momentum equation is

$$\rho u \frac{\partial u}{\partial x} + \rho v \frac{\partial u}{\partial y} = \mu \frac{\partial^2 u}{\partial y^2} .$$

For the case of $Pr = 1$, we clearly have a special solution $h = au + b$, with constants a and b, suitable as the solution if h assumes constant values at $y = 0$ and as $y \to \infty$, i.e., constant wall and free stream temperatures. The heat transfer at the wall is, in such cases

$$\dot{q} = k \frac{\partial T}{\partial y}\bigg|_{y=0} = \mu \frac{\partial h}{\partial y}\bigg|_{y=0} = a\mu \frac{\partial u}{\partial y}\bigg|_{y=0} = a\tau_0$$

or $\dot{q}/\tau_0 = a$, τ_0 being the shear stress at the wall. This is known as "Reynolds' analogy" (of heat transfer and skin friction), ordinarily cast in terms of non-dimensional coefficients.

In the case of compressible fluids, the boundary layer concept can be used to derive a set of simplified equations similar to Eq. (XII.1). Through the variable density, the momentum and energy equations become coupled and must be solved simultaneously, adding much complexity in the solution. Only in special cases may the problem be reduced to an equivalent incompressible one through suitable transformations. At hypersonic speeds, the shock wave, whether detached or not, tends to approach the body surface. Then the inviscid and viscous layers would interact with each other, or even merge together. All these phenomena require considerable finesse in handling. For general three-dimensional bodies in unsteady motion, the theory is yet in an undeveloped stage.

THE SPHEROIDAL METHOD IN THE THEORY OF ARTIFICIAL SATELLITE MOTION

by

J.P. Vinti
National Bureau of Standards
Washington, D.C.

The potential function for the Earth can be written as

$$V = -\frac{\mu}{r}\left[1 - \sum_{n=2}^{\infty}\left(\frac{r_e}{r}\right)^n J_n P_n (\sin\theta)\right] + \text{Tesseral harmonics} \qquad (1)$$

where

θ: latitude

r: geocentric distance

r_e: Earth's equatorial radius.

Most theories for satellite motion employ perturbation methods starting with the unperturbed potential, $V_0 = -\mu/r$. However, because of the oblate shape of the Earth, it is possible to choose a zero-order potential which is a better approximation and at the same time leads to a separable Hamilton-Jacobi equation.

The motivation is based upon the fact that the Hamilton-Jacobi equation is separable in oblate-spheroidal coordinates. These coordinates can be defined by the relations

$$x + iy = r\cos\theta\,\epsilon^{i\theta} = \left[(\rho^2+c^2)(1-\eta^2)\right]^{\frac{1}{2}}\epsilon^{i\theta} \qquad (2)$$

$$Z = r\sin\theta = \rho\eta$$

$$\xi = \rho/c\ ,$$

where for large values of r

$$\rho \sim r \qquad (3)$$

$$\eta \sim \sin\theta\ .$$

The coordinates are ρ, η, γ; γ is the longitude of right ascencion.

It can be demonstrated that the Hamilton-Jacobi equation is separable in ρ, η, γ if and only if

$$V = \frac{f(\xi) + g(\eta)}{\xi^2 + \eta^2} . \tag{4}$$

One then looks for the most general functions f and g which satisfy the following form of the Laplace equation and do not lead to singularities on the ξ-axis:

$$\frac{\partial}{\partial \xi}\left[(\xi^2 + 1)\frac{\partial V}{\partial \xi}\right] + \frac{\partial}{\partial \eta}\left[(1 - \eta^2)\frac{\partial V}{\partial \eta}\right] = 0. \tag{5}$$

It can be shown that the most general form of V that satisfies these conditions is

$$V = \frac{b_0 \xi}{\xi^2+\eta^2} + \frac{b_1 \eta}{\xi^2+\eta^2} = b_0 \,\mathrm{Re}\,(\xi+i\eta)^{-1} + b_1 \mathrm{Jm}\,(\xi+i\eta)^{-1} . \tag{6}$$

The potential V can be expanded into spherical harmonics in the following manner:

$$(\xi+i\eta)^2 = \xi^2 - \eta^2 + 2i\xi\eta = \frac{r^2}{c^2}\left(1 + \frac{2ic}{r}\sin\theta - \frac{c^2}{r}\right) \tag{7}$$

$$(\xi+i\eta)^{-1} = \frac{c}{r}\left(1 + \frac{2ic}{r}\sin\theta - \frac{c^2}{r^2}\right)^{-\frac{1}{2}}$$

$$= \frac{c}{r}\sum_{n=0}^{\infty}\left(\frac{-ic}{r}\right)^n Pn\,(\sin\theta) . \tag{8}$$

Therefore, we can write

$$V = \frac{b_0 c}{r}\left[1 - \frac{c^2}{r^2}P_2 + \frac{c^4}{r^4}P_4 + \ldots\right] \frac{-b_1 c^2}{r^2}\left[P_1 - \frac{c^2}{r^2}P_3 + \ldots\right] . \tag{9}$$

For large r, we must choose $b_0 c = -\mu$; to satisfy (1), we require $c^2 = r_e^2 J_2$. It can be demonstrated that

$$b_1 c^2 = \mu r_e J_1 = -\mu \bar{\xi} \tag{10}$$

$\bar{\xi}$ = coordinate of mass center.

Thus,

$$V = \frac{\mu}{r}\left[1 - \frac{r_e^2}{r^2} J_2 P_2 - \frac{r_e^4}{r^4} J_4 P_4 + \ldots\right] \tag{11}$$

where

$$\begin{aligned} J_4 &= -J_2^2 \\ J_6 &= J_2^3 \\ J_8 &= -J_2^4, \text{ etc.,} \end{aligned} \tag{12}$$

and all odd J's vanish.

For

$$V = \frac{-\mu\rho}{\rho^2 + c^2\eta^2}, \tag{13}$$

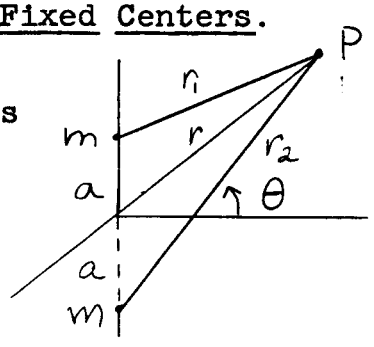
the singularities at $\rho/c = \eta = 0$ are focal circles.

Connection with Problem of Two Fixed Centers.

The potential at point P is

$$-\frac{V}{G} = \frac{m}{r_1} + \frac{m}{r_2}, \tag{14}$$

where



$$r_1^2 = r^2 + a^2 - 2ar \sin\theta$$
$$r_2^2 = r^2 + a^2 + 2ar \sin\theta \,. \tag{15}$$

These lead to the expansions

$$\frac{1}{r_1} = \frac{m}{r_1} \sum_{\ell=0}^{\infty} \left(\frac{a}{r}\right)^{\ell} P_{\ell}(\sin\theta) = \frac{2m}{r_1} \sum_{\ell=0}^{\infty} \left(\frac{a}{r}\right)^{2n} P_{2n}(\sin\theta) \tag{16}$$

$$\frac{1}{r_2} = \frac{2m}{r_2} \sum_{\ell=0}^{\infty} \left(-\frac{a}{r}\right)^{2n} P_{2n}(\sin\theta) \,. \tag{17}$$

Let

$$\begin{aligned} a &= c\sqrt{-1} \\ 2m &= M \\ 2Gm &= \mu \,. \end{aligned} \tag{18}$$

The potential in (14) can then be written as

$$V = -\frac{\mu}{r}\left[1 - \frac{c^2}{r^2} P_2 + \frac{c^4}{r^4} P_4\right] \,. \tag{19}$$

When $c^2 = r_e^2 J_2$, we have the same result as (11), but we don't know that this is the most general potential that results in separability.

The coordinates for the two centers are

$$\xi = \frac{r_2 + r_1}{2a}$$
$$\eta = \frac{r_2 - r_1}{2a} \,, \tag{20}$$

or

$$r_2 = a(\xi + \eta)$$
$$r_1 = a(\xi - \eta) \,. \tag{21}$$

If $a = ic$, one can verify that

$$Z = c\xi\eta$$
$$x + iy = c\left[(\xi^2 + 1)(1 - \eta^2)\right]^{\frac{1}{2}} e^{i\theta} . \qquad (22)$$

One can ask in what sense the two potentials are identical. In the problem of two centers, one can distribute mass uniformly between the two singularities. For the more general potential, distributing the mass uniformly over the focal circle does not produce the same field.

Accuracy of Vinti Potential.

The geoid is defined by the relation

$$\frac{\mu}{r}\left[1 - \left(\frac{r_e}{r}\right)^2 J_2 P_2 + \epsilon\right] + \frac{1}{2}\omega^2 r_e^2 \cos^2\theta = \text{constant}, \qquad (23)$$
$$\epsilon = O(J_2^2).$$

Thus, through terms of $O(J_2^2)$, the geoid is

$$\frac{\mu(1+\epsilon)}{r}\left[1 - \left(\frac{r_e}{r}\right)^2 J_2 P_2\right] + \frac{1}{2}\omega^2 r_e^2 \cos^2\theta = \text{constant}. \qquad (24)$$

Neglecting any harmonics of coefficient ϵ beyond J_2 produces a fractional error, ϵ.

Eg. $J_3 = 2.3 \times 10^{-6}$.

Since the separable potential is accurate through J_2, the error of the corresponding geoid is less than

$$\left[|J_2| + |J_4 + J_2|^2 + |J_3| + |J_6 - J_2^3| + \ldots |J_{12}|\right] \cdot r_e, \qquad (25)$$

this value probably being less than 120 feet.

Separation of Hamilton-Jacobi Equation.

$$H = \sum p^2 - \frac{\mu \rho}{\rho^2 + c^2 \eta^2} . \tag{26}$$

Let

$$S = S_1(\rho) + S_2(\eta) + S_3(\varphi) . \tag{27}$$

The separation constants are α_1 = energy, α_2, α_3 = Z-component of angular momentum.

The isolating integrals are

$$P_\rho = \frac{\partial S_1}{\partial \rho} , \quad P_\eta = \frac{\partial S_2}{\partial \eta} , \quad P_3 = \frac{\partial S}{\partial \phi} = \alpha_3 . \tag{28}$$

One finds integrals of the form

$$\begin{aligned} S_1 &= \int \pm [\,] \, d\beta \\ S_2 &= \int \pm [\,] \, d\eta \\ S_3 &= \alpha_3 \phi . \end{aligned} \tag{29}$$

$$t + \beta_1 = \frac{\partial S}{\partial \alpha_1} = \int_{\rho_1}^{\rho} \pm [\,] \, d\rho + \int_0^{\eta_0} \pm [\,] \, d\eta \tag{30}$$

$$\beta_2 = \frac{\partial S}{\partial \alpha_2} = \int_{\rho_1}^{\rho} \pm [\,] \, d\rho + \int_0^{\eta_0} \pm [\,] \, d\eta \tag{31}$$

$$\beta_3 = \frac{\partial S}{\partial \alpha_3} = \varphi + \int_{\rho_1}^{\rho} \pm [\,] \, d\rho + \int_0^{\eta_0} \pm [\,] \, d\eta . \tag{32}$$

$$G(\eta) = -2\alpha_1 c^2(\eta_0^2 - \eta^2)(\eta_2^2 - \eta^2); \quad (33)$$

motion confined between two hyperboloids:

$$-1 \leq \eta_0 \leq \eta \leq \eta_1 \leq 1$$

$$\eta_2^2 >> 1 \geq \eta_0^2 .$$

$$F(\rho) = (-2\alpha_1)(\rho - \rho_1)(\rho_2 - \rho)(\rho^2 + A\rho + B); \quad (34)$$

motion for negative energy, $\alpha_1 < 0$, is confined between two spheroids: $\rho_1 \leq \rho \leq \rho_2$.

η_0, η_2, ρ_1, ρ_2, A, B are all functions of α_1, α_2, α_3 and some function of the initial conditions.

Following Izak, new variables are introduced:

$$a = \frac{1}{2}(\rho_1 + \rho_2)$$

$$c = \frac{\rho_2 - \rho_1}{\rho_2 + \rho_1} \quad (35)$$

$$I = (\operatorname{sgn} \alpha_3) \sin^{-1} \eta_0 .$$

The real difficulty lies in inverting (30) and (31) to solve for $\rho + \eta$.

This is accomplished by first introducing the uniforming variables E, V, Ψ, X defined by the relations

$$\rho = a(1 - e \cos E) = \frac{a(1-e^2)}{1+e \cos r} \quad (36)$$

$$\eta = \eta_0 \sin\Psi$$

$$\sin X = \frac{\cos\Psi}{[1 - \eta_0^2 \sin^2\Psi]^{\frac{1}{2}}}$$

$$\cos X = \frac{\left[1-\eta_0^2\right]^{\frac{1}{2}} \sin\psi}{\left[1-\eta_0^2\sin^2\psi\right]^{\frac{1}{2}}} .$$

Assume the expansions

$$\begin{aligned} E &= E_s + E_0 + E_1 + E_2 \\ V &= V_s + V_0 + V_1 + V_2 \\ X &= X_s + X_0 + X_1 + X_2 , \end{aligned} \tag{37}$$

where the "s" subscript denotes the secular part. Let

$$E_s = V_s = M_s = \text{secular part of mean anomaly.} \tag{38}$$

One finds

$$\dot{M}_s = 2\pi\nu_1 , \quad \nu_1 = \frac{\partial\alpha_1}{\partial J_1} \tag{39}$$

where $J_i = \oint p_i dq_i$ is the action variable;

$$\dot{\varphi} = 2\pi\nu_2 , \quad \nu_2 = \frac{\partial\alpha_2}{\partial J_2} . \tag{40}$$

Including the periodic terms of order J_2,

$$M_s + E_0 - e' \sin (M_s + M_0) = M_s \tag{41}$$

$$e' = \frac{a}{a_0} e$$

$$a_0 = \frac{-\mu}{2\alpha_1}$$

$$a = a_0 + O(J_2) .$$

V is then found from the relations

$$\cos V = \frac{\cos E}{1-e\cos E}$$

$$\sin V = \frac{(1-e^2)^{\frac{1}{2}}\sin E}{1-e\cos E} \quad . \tag{42}$$

By including periodic terms of order J_2, one can also find relations for X_0 in terms of V_0.

The terms left out of the gravity potential are

$$\Delta V = \mu \frac{r_e^3}{r^4} J_3 P_3 (\sin\theta) + \frac{\mu r_e^4}{r^5} (J_4 + J_2^2) P_4 (\sin\theta) + \ldots . \tag{43}$$

Effects of 99.5 Percent Aspherical Deviations.

	Secular	Short Period	Long Period
Kozai	J_2^3	J_2^2	Doesn't exist
Vinti	J_2^∞	J_2^2	Doesn't exist

Effects of Remaining 0.5 Percent Aspherical Deviations.

				Algorithm
Kozai	J_2^3	J_2^2	J_2^2	Very long
Vinti	J_2^2	J_2^2	J_2	Long

Physical Experiments in Zero g Laboratories

J. P. Vinti
National Bureau of Standards
Washington, D. C.

Lecture notes prepared by Ralph Deutsch.

Reference: J. Research NBS, 67 c, July - Sept. 1963.

1. Forces Acting on Satellite

The principal forces acting on an artificial earth satellite are:

a) gravitational forces including zonal and tesseral harmonics

b) lunar - solar perturbations

c) spin orbit interactions

d) nongravitational forces including

(i) atmospheric drag

(ii) meteoritic impact

(iii) solar radiation pressure

(iv) charge in electric field; induced electric dipole in non - uniform electric field

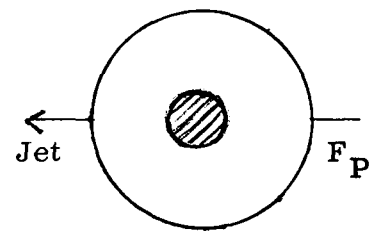
(v) charged body induced magnetic moment; moving relative to the earth's magnetic field

(vi) ferromagnetic currents

(vii) induced currents in satellite producing eddy current effects

In principle, all the non - gravity forces can be neutralized in the satellite by using precisely controlled jet motors in conjunction with a suitable instrument which indicates the presence and direction of any non - neutralized gravity force.

2. Unmanned Double Satellite



The nongravitational forces can be neutralized by the jet, illustrated in the figure. For a spherical satellite, the zero - gravity condition is maintained by keeping the sensing element in the center of the satellite by using a servo - mechanism to control the jets. At lower altitudes, the atmospheric drag is the most important of the nongravitational forces. Thus the amount of jet thrust required to keep the test object centered is a measure of the drag force.

3. Manned Space Capsule

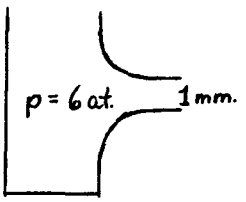
a) The test instrument can be a sphere containing a small ball. The sphere must be rigidly attached to the capsule, be electromagnetically shielded from stray fields, and have an opening for observing and measuring the displacement of the ball. The external capsule jets might be under control of an astronaut who would keep the test object centered at the center of mass.

b) Example: Consider a Mercury - type capsule in a 200 km. orbit.

$$\text{mass} = 10^6 \text{ gm.}$$
$$\text{drag} = 6 \times 10^4 \text{ dyne}$$
$$= 1.6 \text{ oz.}$$

At 300 km., the drag is 6×10^3 dyne.

c) Compressed Air Jet:



One possible type of control would be a compressed air jet illustrated in the figure. Suppose that the orifice had a diameter of 1 mm. and the air is at 6 atmospheres pressure. The thrust generated by the jet is

$$T = \frac{50 \text{ kg. force sec.}}{\text{kg. mass}}$$

To produce the required 6×10^7 dynes corresponding to the height of 200 km., the air reservoir would require 6.8 kg. per circular orbit. If the orbit's eccentricity is e = 0.03, 1.6 kg. of air would be required per orbit.

On the other hand, increasing the orbit height to 300 km. reduces the preceding air capacity figures by a factor of ten. However, for long flights of many orbits, the required weight of the reaction gas would be excessive.

d) Chemical jets:

Microchemical jets are being developed which will be far more efficient than the simple compressed air jets. The microchemical jets have thrust capabilities of the order of

$$T = \frac{300 \text{ kg. force sec.}}{\text{kg. mass}} .$$

These produce 6 times the thrust of the air jets at 1/6 of the mass of the required jet material. For example

	Altitude	Weight per Orbit
Circular	200	1 kg.
Circular	300	0.1
e = 0.03	200	250 gm.
e = 0.03	300	25 gm.

3. Gravitational Field Inside a Capsule and Relative to It

Let x, y, z be a set of inertial coordinates.

Then

$$M\ddot{\mathbf{R}} = \mathbf{F}_G + \mathbf{F}_D,$$

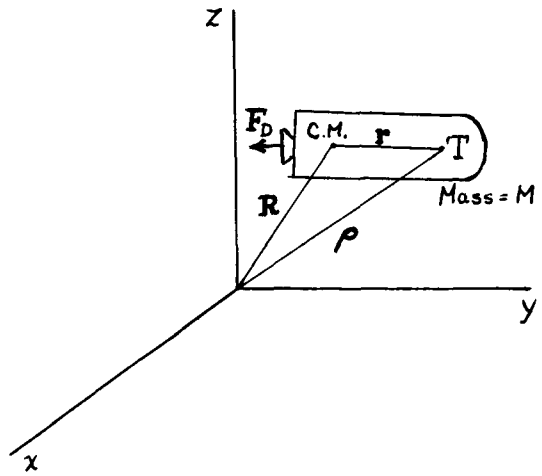
$\mathbf{F}_G$ = gravity force

$\mathbf{F}_D$ = drag force

$\mathbf{R}$ = vector to the center of mass; C.M.

For the test object, T,

$\ddot{\mathbf{R}} + \ddot{\mathbf{r}} = \mathbf{f}$.



Therefore,

$$\ddot{\mathbf{R}} = \frac{\mathbf{F}_G}{M} + \frac{\mathbf{F}_D}{M}$$

and

$$\ddot{\mathbf{r}} = \mathbf{f} - \ddot{\mathbf{R}} = \mathbf{f} - \frac{\mathbf{F}_G}{M} - \frac{\mathbf{F}_D}{M} ,$$

is the gravitational field acting on the test object relative to an inertially oriented capsule. If the capsule is not inertially oriented, one has to add the apparent forces, $-2\omega \times \dot{\mathbf{r}}$, $\omega \times (\omega \times \mathbf{r})$ and the $\dot{\omega}$ force.

For a spherically symmetric capsule

$$\mathbf{f}_{C.M.} = \frac{\mathbf{F}_G}{M} , \text{ rigourously.}$$

If the capsule is not spherically symmetric, then the preceding relation is an approximation. The gravitational acceleration relative to the capsule is

$$\mathbf{g} = \ddot{\mathbf{r}} = \mathbf{f} - \mathbf{f}_{C.M.} - \frac{\mathbf{F}_D}{M} .$$

For small capsules not at very high altitudes,

$$|\mathbf{f} - \mathbf{f}_{C.M.}| << \frac{\mathbf{F}_D}{M} .$$

Thus, as an approximation

$$\mathbf{g} = -\frac{\mathbf{F}_D}{M} .$$

4. Determination of Drag by Measurements Inside Capsule

Let

$$a = \left| \frac{\mathbf{F}_D}{M} \right| .$$

Then if the test object starts from rest at or near the C.M., in a time t it will travel a distance

$$s = 1/2\ a t^2 .$$

It is assumed that the test object is contained in a housing that is evacuated and well shielded from electromagnetic fields.

For example, for a Mercury-type capsule at 200 km.,

$|F_D| = 6 \times 10^4$ dynes

$M = 10^6$ gm.

$a = 6 \times 10^{-2}$ cm./sec.^{2}

In 10 seconds, the test object will have moved

$s = 1/2\ (6 \times 10^{-2})\ 100 = 3$ cm.

At 300 km., one finds $s = 3$ mm.

There are a few objections to the proposed measurement technique.

(a) Apparent forces: Centrifugal forces entirely negligible and produce errors of about 1/50 a.

(b) Accelerations produced by body motions of astronaut: The main irremovable effect is that caused by heart beats which can produce instantaneous accelerations of about 10 a. These accelerations vary very rapidly and tend to smooth out in the inertia of the mechanical systems.

Ballistocardiography tests have indicated peak displacements from a mean position of 0.03 mm. for a human subject coupled to a capsule. The corresponding displacement of the test object would be only 0.003 mm. compared with $s = 3$ mm. at a height of 300 km. produced by drag.

5. Determination of Atmospheric Density

If the drag force F_D is measured, the atmospheric density can be computed from

$$F_D = 1/2\ C_D A \rho v^2 .$$

For most capsules $C_D \approx 2.3$, v^2 is known from the orbit and A is known from the design of the capsule.

For non-spherical satellite, A can be maintained at a constant value during the measurement of atmospheric density by keeping an axis of the capsule parallel to the velocity vector. Such control can be accomplished by control jets governed by a pair of static accelerometers.

6. Determination of Perigee Passage

At perigee, the velocity is

$1/2\ v^2 = E + \mu/r$.

v^2 has a maximum at perigee; moreover, the air density ρ is also a maximum at perigee. Even for low eccentricity orbits such as $e = 0.03$, the maximum of ρ is very sharp. Therefore, by continuously monitoring the drag force, the astronaut can determine perigee passage and by the use of a clock can determine the time of perigee passage.

7. Simultaneous Gravity Orbit and Constant Monitoring of Drag

By intermittently operating the control jets, the astronaut could measure the drag and return the test object to the center of the housing at each test so as never to permit any collisions within the accuracy of the observations. Note that under these conditions, the orbit of the capsule is then the same as that of the test object and is thus gravitational.

Thus we have a technique of simultaneously producing a gravity orbit and determining the time of perigee passage in each orbit. With proper cooperation from ground observers, one can also determine the position of perigee on each orbit. This type of controlled gravity orbit might be very useful for geodetic purposes, both to determine the potential coefficients J_n and to determine station location errors.

8. Re-entry Meter

Consider a case in which $a = 10^{-3}$ g. Then in 5 seconds the test object would move

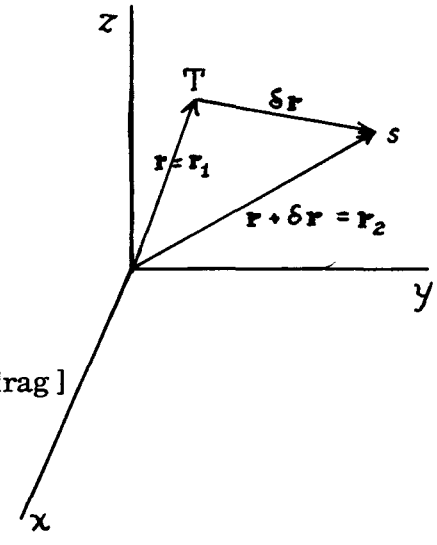
$s = 1/2\ at^2 = 1/2 \times 1 \times 25 = 12.5$ cm.

Thus the test object in its evacuated housing combined with a stop watch is a sensitive g-meter which would be simple and almost perfectly fail-proof.

9. Tacking the Drag on to a Gravitational Orbit

As the astronaut sits in the capsule, watching the test object drift along, the motion that he observes is the difference between the true orbit of the capsule and the gravitational orbit of a particle that starts out with the same position and velocity vectors. This can be explained by considering the figure.

Let $\mathbf{r}$ be the position vector of a very heavy particle, and $\mathbf{r} + \delta\mathbf{r}$ be the position vector of a satellite which starts out with the same initial conditions as the heavy particle.



Let:

(i) $\mathbf{r} = \mathbf{r}$ [accurate gravity field, no drag]

(ii) $\mathbf{r} + \delta\mathbf{r} = \mathbf{r}_1$ [approximate gravity field, with drag]

$= \mathbf{r}_1$ [$v = -\frac{\mu}{r}\left(1 - \frac{r_e^2}{r^2} J_2 P_2\right)$ + drag]

(iii) $\mathbf{r}_2 = \mathbf{r}_2$ [approximate gravity field, no drag]

$\delta\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$.

For short time intervals, J_2 does not have to be included in $\mathbf{r}_1$, because the line of apsides moves slowly (4°/day). Thus J_2 can be neglected for times less than a day. In any case, one doesn't need any terms beyond J_2, drag, $J_2 \times$ drag, in the solution for $\mathbf{r}_1$. In particular no terms of the order J_2^2 are required.

It can be concluded that in the above fashion, the problem of drag can be separated from the problem of the accurate gravitational orbit. To find the effect of drag, solve the problems (ii) and (iii), then the difference $\delta\mathbf{r}$, and add to $\mathbf{r}$.

It should be observed that this method will not work for very long time intervals. For very long intervals one cannot depend upon knowledge of the air density function $\rho(r, \theta)$. The air density function is so poorly known that one will have to redetermine the orbital elements long before J_2^2, etc., become necessary in the expression for $\mathbf{r}_1$. In other words, the problem of atmospheric drag is not sufficiently well modeled mathematically to warrant a very accurate mathematical treatment.

10. Zero g Laboratories - Achieved with jets and floating test object

(a) Point of zero g characterized by $f = F_G/M$. It is at the center for capsule with spherical symmetry

(b) In other cases, point of zero g may or may not exist. It probably would not exist in a space laboratory having the shape of a torus. Usually, if it exists, it will be close to the C.M. Let us assume that the departures from zero g at the C.M. are small.

(c) Possible experiments in zero g laboratories: free top, liquid shapes and motions, all experiments impossible on earth because of all-pervasive gravity field.

(d) Experiments in dynamical astronomy, with gravitational or electrostatic forces, or both. In this lecture we will only discuss the determination of the constant G.

(e) It has been noted that the period of a close satellite of the moon is the same as a close earth satellite. This implies that the mean densities must be equal, or

$$G\frac{4}{3}\rho a^3 = n^2 a^3 .$$

Thus, a marble travelling in a close circular orbit around a sphere whose mean density approximates that of the earth would have a period T = 84 minutes. Or more accurately,

$$G(M+m) = \left(\frac{2\pi}{T}\right)^3 a^3 ,$$

where a is the semi-major axis of the orbit which need not be a perfect circle.

The large sphere used to model the earth could float at the C.M. of the laboratory; it would be kept at this point by external jets which are controlled to maintain a gravity orbit. Thus if this sphere were enclosed in an evacuated spherical housing with electromagnetic shielding and an observation port, then the sphere itself could serve as the test object for maintaining the gravity orbit. A marble could then be placed in orbit about the sphere by accurately controlling the initial conditions to insert the marble in a close, nearly circular orbit. If M, m, and T are measured and a can be determined with sufficient precision, then the preceding relation can be used to compute G.

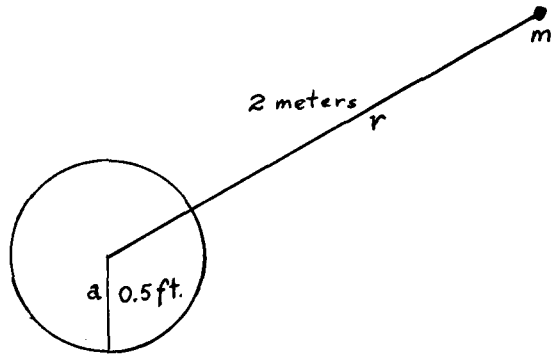
11. Lack of Spherical Symmetry

Consider the acceleration of the mass

$$f = \frac{G\frac{4\pi}{3}a^3}{a^2} = G\frac{4\pi}{3}\rho a .$$

For a perfect sphere of radius r_e,

$$g_e = G\frac{4\pi}{3}r_e^3 \cdot \frac{1}{r_e^2} = G\frac{4\pi}{3}\rho r_e .$$



thus

$$f = \frac{a}{r_e}g_e = \frac{1/2}{2\times 10^7}\times 10^3 = \frac{1}{4}\times 10^{-4} \text{ cm./sec.}^2$$

$$\approx \frac{1}{40} \text{ milligal} .$$

Suppose a perturbing mass, $m = 10^5$ gm. is placed at a distance r = 200 cm.

from the C.M. the change in acceleration would be

$$\delta f = \frac{Gm}{r^2} - \frac{Gm}{(r-a)^2} \approx \frac{2Gm}{r^3} a \ .$$

For example,

$G = 6.7 \times 10^{-8}$ cgs units

$m = 10^5$ gm.

$r = 200$ cm.

$a = 15$ cm. ;

$\Delta f = 25 \times 10^{-9}$ cgs units;

and $\frac{\Delta f}{f} = \frac{25 \times 10^{-9}}{1/4 \times 10^{-4}} = 10^{-3}$, which is too high a value.

If the asymmetry can be reduced to 10^3 gm. at 200 cm., then we obtain the better case of

$$\frac{\Delta f}{f} = 10^{-5} \ .$$

There are several difficulties that must be considered in the proposed experiment for the determination of G .

(i) Lack of spherical symmetry in surroundings. The effect of nonspherical environments can be minimized by floating the large sphere which is accelerated about as much as in the marble. The effect of nonspherical environment is equivalent to the lunar - solar perturbations of an earth satellite, $\sim 1/r^3$. These effects might be further diminished by using shims and having personnel move countermasses to balance their own movements in the laboratory.

(ii) Electrostatic charges on sphere and marble. It is impossible to completely shield the experiment electrostatically from charge because of the presence of high velocity cosmic rays. It will be necessary to avoid the Van Allen belts for these experiments.

(iii) The control sphere used to model the earth will have unavoidable lack of complete symmetry. Thus the problem is more complicated and will have to be treated as a nonspherical earth and must be represented by zonal harmonics. Probably the dominant terms will be J_1, J_2, J_{22}.

J_1 can be found by determining the C.M. of the sphere using a compound pendulum. This determination can be accomplished on the ground before flight. J_2 and J_{22} can be determined by measuring the principal moments of inertia A, B, C on the ground by means of a torsion pendulum. This is done by determining the moments of inertia about the various axes, then

$$J_2 = \frac{C - \frac{1}{2}(A+B)}{ma^2}$$

$$J_{22} \approx \frac{1}{2}(A-B) .$$

The effects of the nonspherical sphere can then be accounted for by applying an appropriate theory for satellite orbits about an oblate earth.